AD___________________

Award Number: W81XWH-05-1-0293

TITLE: Computer Aided Detection of Breast Masses in Digital Tomosynthesis

PRINCIPAL INVESTIGATOR: Swatee Singh
Joseph Lo Ph.D.

CONTRACTING ORGANIZATION: Duke University
Durham, NC 27710

REPORT DATE: June 2008

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*
01-06-2008

2. REPORT TYPE
Annual Summary

3. DATES COVERED *(From - To)*
1 Jun 2005 – 31 May 2008

4. TITLE AND SUBTITLE

Computer Aided Detection of Breast Masses in Digital Tomosynthesis

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-05-1-0293

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Swatee Singh and Joseph Lo Ph.D.

E-Mail: swatee.singh@duke.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Duke University
Durham, NC 27710

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES – Original contains colored plates: ALL DTIC reproductions will be in black and white.

14. ABSTRACT

e purpose of t is study as to propose and implement a computer aided detection (CADe) tool for breast tomosynt esis using data from 100 uman subject cases. nli e traditional CADe algorit ms in ic t e second stage P reduction is done ia feature e traction and analysis, instead information t eory principles ere used it mutual information as a similarity metric. ree sc emes, A, and C, ere proposed and t ey differed in t e composition of t eir no ledge base of regions of interest (Is). Sc eme A s no ledge base as comprised of all t e mass and P Is generated by t e first stage of t e algorit m ile sc eme contained information from mass Is and randomly e tracted normal Is and sc eme C contained masses, Ps and normal Is. e results indicated t at t e best o erall system performance as 85 sensiti ity it 2.4 Ps per breast olume for sc eme A, 3.6 Ps per breast olume for sc eme and 3 Ps per breast olume for sc eme C.

15. SUBJECT TERMS
computer aided detection, digital mammography, sub-region hotelling observer, digital tomosynthesis, multi-slice CAD algorithms, biopsy

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	131	**19b. TELEPHONE NUMBER** *(include area code)*

TABLE OF CONTENTS

Introduction……………………………………………………………….……..…..... 4

Body……………………………………………………………………………..……… 4

Key Research Accomplishments……………………………………….………. 14

Reportable Outcomes……………………………………………………………….. 15

Conclusions…………………………………………………………………………….. 16

References……………………………………………………….………………….... 16

Appendices……………………………………………………….………………..…. 17

INTRODUCTION

Mammography is currently the most effective early-detection tool for breast cancer screening. To provide a reliable and efficient second-reader to aid breast-imaging radiologists, recent research has been directed towards developing computer-aided detection (CAD) tools for mammography. Although these tools have shown promise in identifying calcifications, detecting masses has proven relatively more difficult primarily due to presence of dense overlying tissue in a mammogram. Breast tomosynthesis has the potential to improve detection and characterization of breast masses by removing overlapping dense fibroglandular tissue. These systems provide 3D slice images from a modified full field digital mammography system which acquires a limited-angle cone beam CT scan under mammography positioning.

The goal of tomosynthesis to provide 3D information at comparable dose, resolution, and patient throughput to mammography, and with lower cost and hardware requirements compared to alternatives such as breast Computed Tomography or breast Magnetic Resonance Imaging. However, with tomosynthesis, instead of the traditional 4 mammography views per case, the radiologist must interpret a large volume of data per breast volume. Given this constraint, the role of CAD is especially important in breast tomosynthesis. If this modality is ever intended to replace mammography as a screening tool, then a CAD algorithm that presents the radiologist with initial cues could potentially become indispensable to maintain current clinical workflow. In fact, investigators in CT colonography have already begun to show that CAD can potentially ease radiologist workflow with large 3D datasets.

As part of our investigation for this grant, we have built a highly sensitive and highly specific CAD scheme for tomosynthesis, incorporating unique preprocessing techniques and advanced decision theory methods. This CAD scheme is expected to detect masses and improve the performance of radiologists attempting to sift through ~50- 80 reconstructed slices of a single breast view. Regardless of whether we choose to work with projection images initially or entirely in the reconstructed domain, the proposed CAD system has two key components: 1) a highly sensitive mass detector, and 2) statistical models designed to reduce false-positives.

BODY

Task 1. Translate single-slice CAD algorithms to individual, reconstructed tomosynthesis slice images:

1.1. Generate independent training and testing subsets of 160 and 40 cases, respectively, from the available 200 patient cases of biopsy-proven masses

Data was collected from the mentor Dr. Lo’s other funded grants. We have a shortfall in terms of total number of biopsy proven cases that were collected by the end of this grant. The pace of subject accrual had been steady since February 2006 after software/ hardware upgrades and the hiring of a clinical coordinator. However, we have found that despite these efforts, our accrual rate has reached an upper limit resulting from unforeseen practical difficulties of running a human subject trial.

Currently, Duke has acquired tomosynthesis data from a total of 241 subjects. Of these, a total of 230 cases have been read by trained radiologists, and 38 subject scans were found to have true lesions in them.

Our experiences with the aims of this goal have led to an abstract being accepted and presented in RSNA 2006 titled "Breast Tomosynthesis: Initial Clinical Experience with 100 Human Subjects," (reportable outcomes #12)

1.2. Apply 2D mammography CAD technique of Hotelling Observer on tomosynthesis slices

1.2.1. Establish baseline performance of existing computer aided detection algorithms on tomosynthesis slices from 160 patient cases

Work on this task began after the work for task 2.2, the results of which steered research for this specific aim towards featureless false-positive reduction techniques used previously in 2D mammography to be applied to tomosynthesis. Baseline performance was thus established using human subject data from 80 cases and was evaluated using only the projection images. In the false positive reduction stage of the algorithm various metrics were implemented and the results were reported as Receiver Operating Characteristic (ROC) Area Under Curve (AUC) by applying a leave-one-out cross validation scheme on all available ROIs. Our best performing metric was the joint entropy with a classifier AUC of 0.87.

Work for this specific task resulted in a proceedings paper at SPIE, the primary scientific conference for medical imaging in 2007 (reportable outcome #10). Techniques from this work were also instrumental in the submission of another paper in SPIE with our industrial collaborators, Seimens Medical Systems (reportable outcome #11), and other Duke collaborators (reportable outcomes #7, 8 and 9) in 2007 and 2008.

Task 2. Extend CAD algorithms for mass detection interrogating a 3D tomosynthesis volume by studying 3D, multi-slice CAD algorithms for lesion detection and characterization:

2.1. Apply 2D CAD to projection images prior to 3D tomosynthesis reconstruction

Work for this task was reported as completed in last years report. The CAD algorithm implemented in this study can be divided into two major stages - the high-sensitivity, low-specificity stage and the false positive reduction stage. Task 2.1 implemented the high-sensitivity, low-specificity stage that can be further divided into two stages - the initial candidate generation and region of interest (ROI) extraction stages. Initial candidate generation has been implemented in this study via filtering of the projection images using an optimized Difference of Gaussians (DoG) filter. This filtered image undergoes multi-level thresholding to yield initial CAD suspicious locations in the 2-D projection images. Please see figure 1 for a diagrammatic representation. In the report for year 1, we have already displayed reconstructed images using filtered back projection.

Techniques from this specific task resulted in a peer-reviewed paper at Academic Radiology in 2008 (reportable outcome #3).

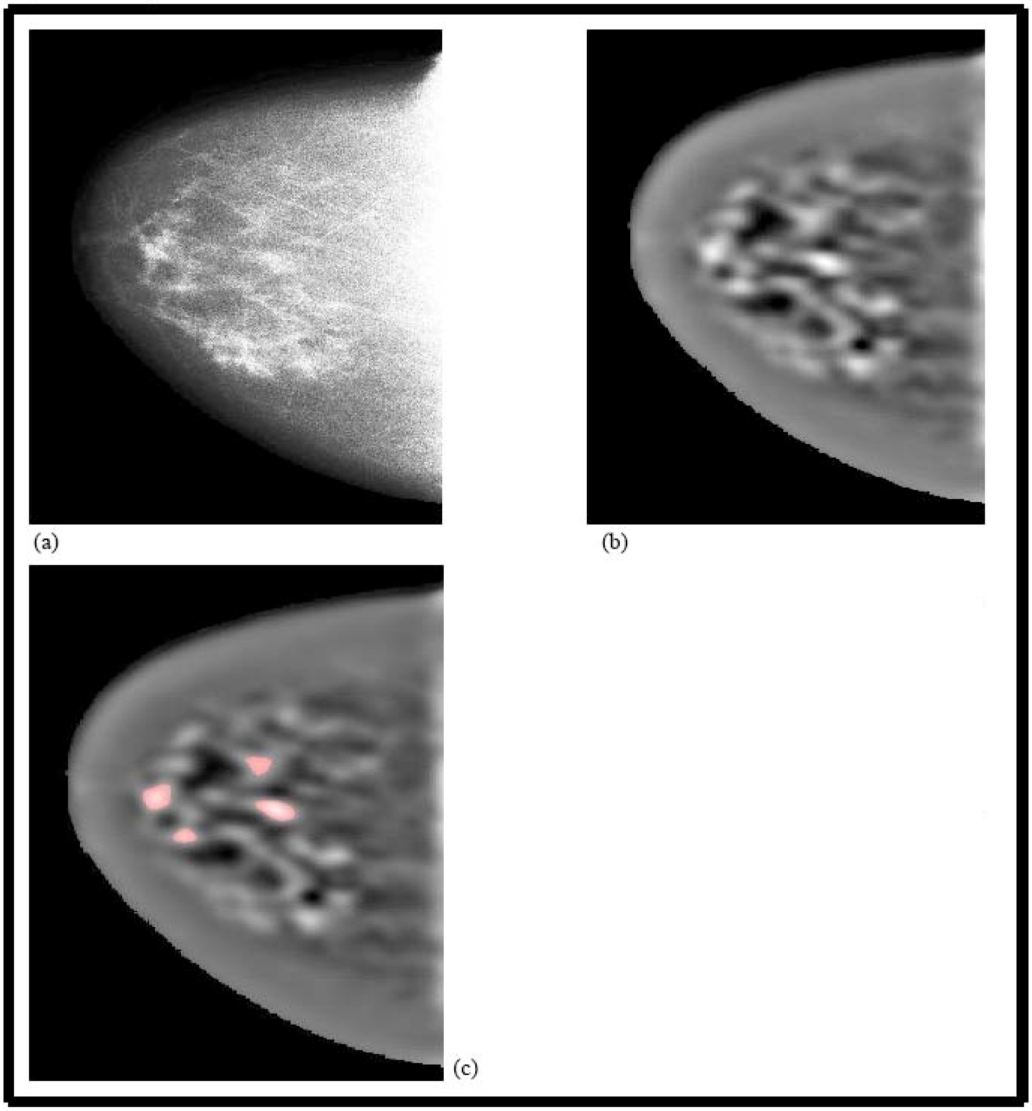


Figure 1: (a) Original middle-projection image of subject 33, LCC view (b) DoG filtered image of image 'a' (c) CAD 2D suspicious location in red overlaid on the DoG filtered image of part 'b'

2.2. Implement a Laguerre-Gauss Channelized Hotelling Observer (LG-CHO) for 3D mass detection

Work on this task began ahead of schedule during the first year because we anticipated that those results might affect our approach for this task. As detailed in the first year's annual report, the Watson filter model significantly outperformed the LG-CHO filters for the task of detection of masses. This finding led us to believe that there were better models for our data than the proposed LG-CHOs. We found that reconstructed tomosynthesis slices from task 2.1 have very little intensity variation across a mass ROI, and hence we anticipated difficulties in the segmentation and feature measurement of these ROIs in the reconstructed domain. As such we chose to use a featureless approach to false positive reduction for this task and decided to work with Information Theoretic CAD (IT-CAD) instead of LG-CHOs to assess image similarity for this task. IT-CAD based similarity assessment relies completely on the statistical properties of the image histograms eliminating the image preprocessing, segmentation, and feature extraction steps. Furthermore, information theoretic similarity measures have the advantage of making no assumptions on the underlying image distributions. This is especially crucial for us given the small number of true lesions in our dataset.

To complete this task, we extended our IT-CAD algorithm to work with ROIs extracted from reconstructed slices instead of projection images. The CAD scheme is comprised of two distinct stages – the 'high-sensitivity, low specificity' stage wherein ROIs are extracted followed by the 'high-sensitivity, high-specificity' stage that uses information theory principles to reduce false positives. Flowcharts of the 2 stages of the algorithm are shown in figures 2 and 3.

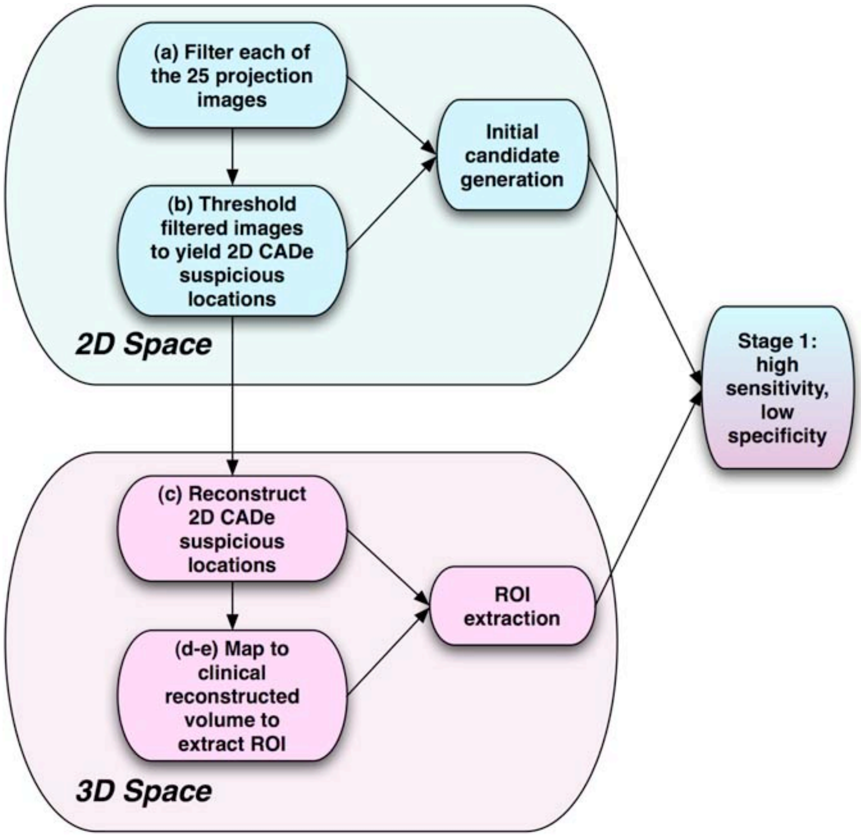

(a) Filter each of the 25 projection images
(b) Threshold filtered images to yield 2D CADe suspicious locations
Initial candidate generation
2D Space
Stage 1: high sensitivity, low specificity
(c) Reconstruct 2D CADe suspicious locations
ROI extraction
(d-e) Map to clinical reconstructed volume to extract ROI
3D Space

Figure 2: Stage 1 – the filtration and ROI extraction or the 'high-sensitivity, low specificity' stage of the CAD algorithm

Once the algorithm had identified initial candidates for mass detection by giving the X, Y and Z location of the centroid of the volume of interest, regions of interests (ROIs) were extracted from the reconstructed breast slice images obtained by filtered backprojection (FBP) which yielded 1 mm thick slices with 85x85 micron pixel pitch. The FP reduction scheme therefore was based upon the same reconstructed image data as used by radiologists. Two sets of ROIs were extracted to assess the effect of information from one versus many slices. In the first set, 256x256 pixel ROIs (22 x 22 mm) centered at the central slice containing the suspicious CAD location were extracted. For the second set, 256x256 ROIs representing the summed slab of 5 slices (5 mm) were extracted. Since lesions typically span multiple reconstructed slices, these two sets investigated whether giving more ‘signal’ to the false positive reduction scheme resulted in an improvement in performance. Use of a slab would also reduce the impact of slight errors of localization in the Z direction.

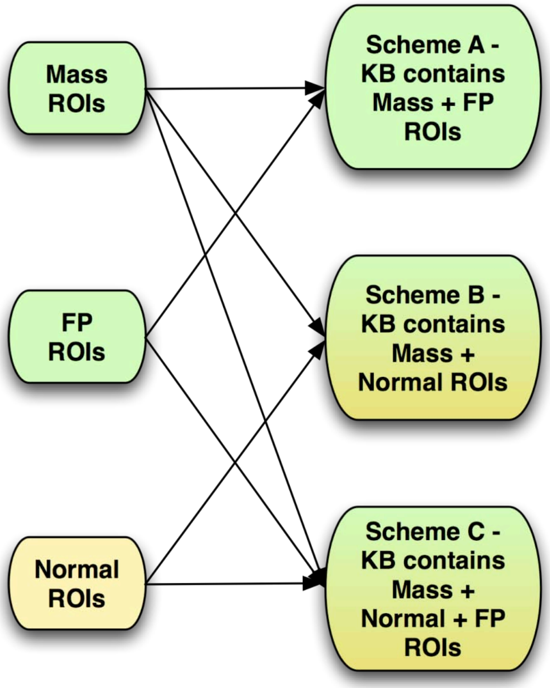

Mass ROIs
FP ROIs
Normal ROIs
Scheme A - KB contains Mass + FP ROIs
Scheme B - KB contains Mass + Normal ROIs
Scheme C - KB contains Mass + Normal + FP ROIs

Figure 3: Composition of Knowledge Base of false positive reduction stage of the CAD algorithm (a) scheme ‘A’ KB composition (b) schemes ‘A’ and ‘B’ composition

Traditionally CAD schemes measure, among others, morphological and texture features of a suspicious location for subsequent false positive reduction using trainable classifiers. This study used mutual information as a similarity metric for false positive reduction that relies completely on the statistical properties of the image histograms and the relationship between pixels of an image. Three schemes were therefore developed for the second stage of the algorithm, as shown in Figure 3. In scheme A, FP reduction was done using a KB containing ROIs from the CAD algorithm’s first stage. These ROIs were either mass ROIs or FPs. In scheme B, the KB contained only mass ROIs and randomly selected normal ROIs from well-separated depths in all the normal cases’ reconstructed volumes. A total of 1390 such normal ROIs were extracted for this study. To access performance of the scheme A classifier, a leave-one-case-out validation scheme was used. Thus, for every ROI that was presented to the system as a query ROI of unknown pathology, all other ROIs generated from that specific subject’s reconstructed volumes were excluded from the KB. For scheme B, all the FPs of the first stage of the algorithm served as queries to the system to assess its specificity. Sensitivity for scheme B was evaluated using a leave-one-case-out sampling scheme on all available ROIs that contained a mass. Thus the system has no knowledge of FP ROIs in its KB and hence the performance is not dependent on the nature of FP lesions generated by the first stage of the algorithm. Finally, scheme C included information from all three sources, (1) masses (2) CAD generated FPs (3) normal breast tissue, combined into a single KB. Analysis was done in a leave-one-case-out manner for this KB as well.

Table 1 presents overall classifier performance for all schemes. As implemented, summing adjacent slices did not improve the classifier performance in a statistically significant way compared to using only the single, central slice ROI for any of the schemes evaluated, either for AUC or partial AUC. Shown in Figure 4 are the ROCs and partial ROCs of just the central slice classifiers of all schemes.

Table 1: Classifier performance for a KB containing mass and FP ROIs (scheme A), mass and normal ROIs (scheme B) and when the KB contains ROIs from all 3 sources – mass, FP and normal ROIs (scheme C). The AUC and pAUC for both the central slice and the sum of adjacent slices and their corresponding p-values for all schemes is shown.

Scheme	Central Slice only		Sum of Adjacent slices		p-value	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
A	0.88 +/- 0.02	0.49 +/- 0.09	0.89 +/- 0.03	0.46 +/- 0.10	0.3	0.2
B	0.86 +/- 0.03	0.41 +/- 0.09	0.89 +/- 0.03	0.36 +/- 0.10	0.5	0.2
C	0.87 +/- 0.02	0.45 +/- 0.09	0.88 +/- 0.03	0.41 +/- 0.10	0.43	0.19

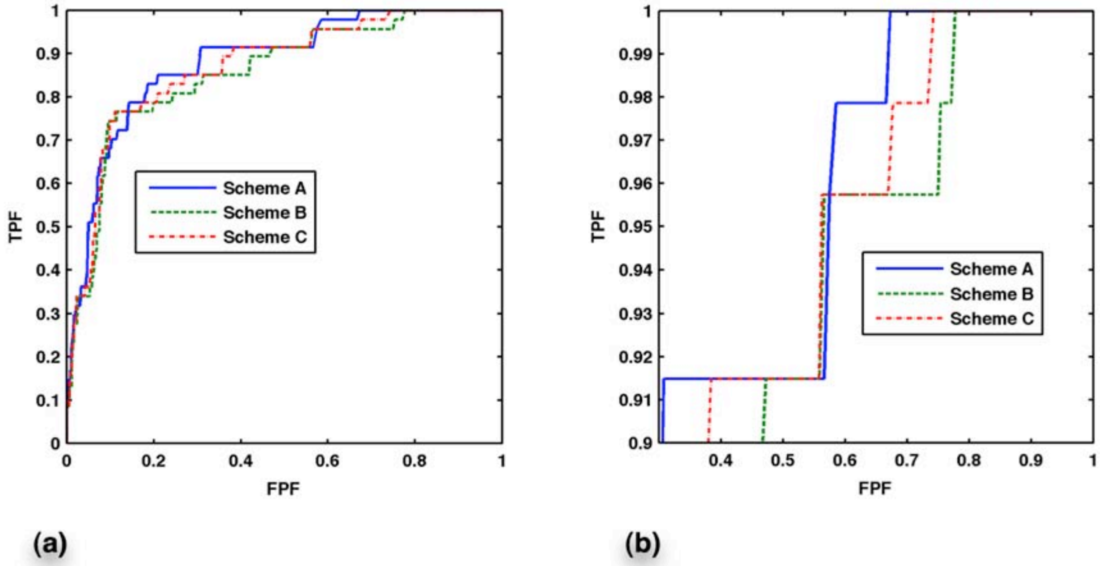


Figure 4: (a) Non-parametric ROC curves of the central slice classifier for schemes A, B, and C (b) Partial ROC curves for sensitivity greater than 0.9 for the three schemes

Sensitivity when plotted as a function of the average FP rate while the decision threshold is varied results in the Free-Response Receiver Operating Characteristic (FROC) curve. Figure 5 shows the system FROCs prior to FP reduction as well as after FP reduction for schemes A, B and C. These were obtained by varying the decision threshold over classifier outputs of the central slice classifiers of the three schemes starting with a threshold set at 91.5% sensitivity. For each scheme, the threshold was then progressively dropped to obtain the entire curve.

Scheme A outperformed others in terms of FPs per breast volume at equivalent sensitivity. At an operating point of 91.5%, scheme A was successfully able to discard 69% of the FPs per breast volume, scheme B correctly eliminated 53% of the FPs per breast volume, and lastly, scheme C was able to correctly discard 62% of the FPs per breast volume. The final performances were a sensitivity of 85% at 2.4 FPs per breast volume, 3.6 FPs per breast volume, and 3 FPs per breast volume for schemes A, B and C respectively. The Jackknife Free-Response Receiver Operating Characteristic (JAFROC) was used to evaluate these FROC curves. None of the differences between the FROC curves of the three schemes studied were statistically significant.

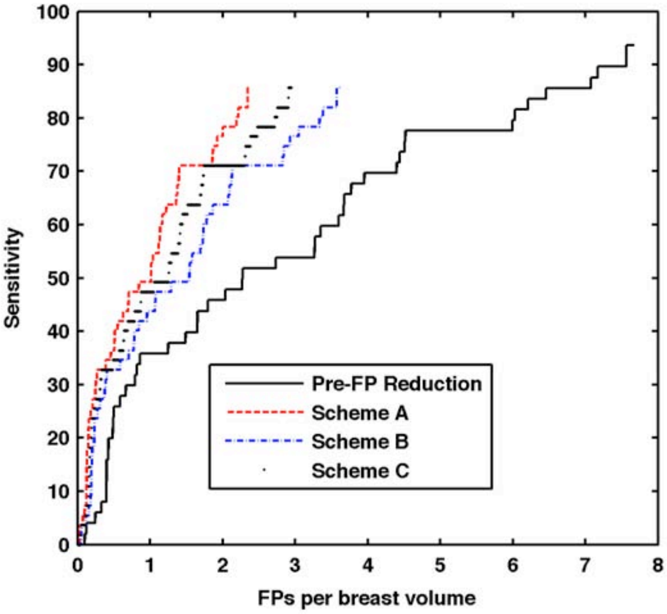


Figure 5: System FROCs. Prior to FP reduction, the system performance was at 93% sensitivity with 7.7 FPs per breast volume. Final system performances for the three schemes are depicted for the central slice classifiers.

A human subject example from subject 122 is shown in Figure 6. While this subject had 5 FPs in total only 2 reconstructed slices containing 1 TP and 2 FPs are shown for illustration purposes. These results were obtained when the CAD algorithm with a scheme A central slice classifier is used while operating at 91.5% sensitivity. After FP reduction, the FP in slice 40 was eliminated, however one FP along with the TP survived in slice 36. This subject had biopsy confirmed cancer.

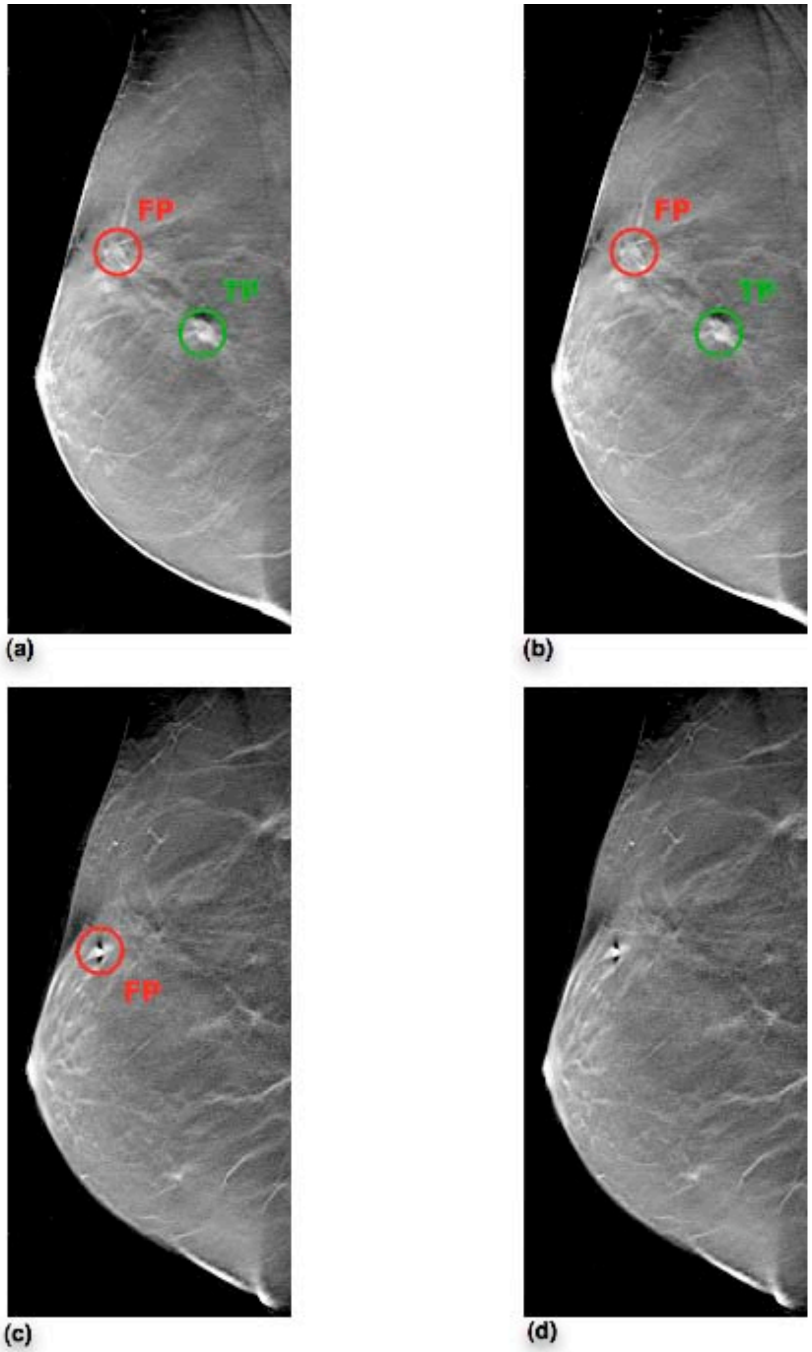


Figure 6: (a) Slice 41 prior to FP reduction (b) Slice 41 after FP reduction (c) Slice 21 prior to FP reduction (d) Slice 21 after FP reduction
Subject 122 had biopsy confirmed carcinoma. While this subject had 6 FPs in total from stage 1 of the CAD algorithm, only reconstructed slices 41 and 21 are shown in this figure for illustration. After setting the threshold for scheme A central slice classifier to operate at 91.5% sensitivity, we are able to eliminate the FP in slice 21. However, the FP in slice 41 survives along with the TP.

A peer-reviewed paper in Medical Physics, a highly respected journal among Medical Physicists, has been accepted based on the results of this task and is currently in press (reportable outcome #1). Work for this specific task also resulted in a proceedings paper at SPIE, the primary scientific conference for medical imaging in 2008 (reportable outcome #6) and in IWDM in 2008 (reportable outcome # 2). Techniques from this work were also instrumental in the submission of 2 other peer-reviewed papers in Medical Physics (reportable outcomes #4 and 5) in conjunction with Duke collaborators in 2007.

Task 3. Evaluate performance of CAD model on independent testing subset of cases (Months 34-36)

Given the small number of cases collected it was not possible to eke out a separate testing set for our algorithm as of the total of 240 human subject data only 38 had lesions in them. However, we evaluated performance of our algorithm using a leave-one-case-out cross validation for all the previously reported tasks.

We also evaluated the performance of the CAD model by evaluating the composition of the knowledge base of our information theory based system. The first stage of the algorithm generates ROIs that are either mass lesions or FPs. Of note here is the imbalance in the number of lesion ROIs when compared to the total number of FPs generated by the first stage. Given this imbalance, it is imperative to explore the effect of knowledge about normal breast parenchyma represented by those FPs. This was studied in two ways. First, an increasing number of FPs was sampled from all FPs available while holding the number of true positives constant, thus decreasing the ratio of mass ROIs in the KB and progressively giving the system more indirect ‘knowledge’ of normal breast parenchyma. The second approach is to provide the system with direct information about normal breast parenchyma via randomly selected normal ROIs instead of suspicious FP regions generated by a CAD algorithm. Since these ROIs were extracted from random locations from within the breast volume there is a potential for some overlap with FPs generated by the first stage of the algorithm. Varying the number of mass ROIs in the knowledge base can also change composition of the knowledge base. However, given that our database consists of a limited number of mass ROIs, its effect was not studied in this experiment.

Scheme A - effect of FP ROIs in the KB

Scheme A seeks to differentiate between a mass and a FP query. A plot of the ROC AUC as a function of increasing number of FPs is presented in Figure 7, where the x-axis shows number of FPs as multiples of the total number of mass ROIs while using the scheme A classifier. The error bars are obtained by simple random sampling from all the available FPs of the first stage. 20 subsets of the FP ROIs were generated for each data point on the graph. Each subset was selected without replacement after randomization between subsets. When the sum of adjacent slices were used, as the number of FPs was increased the performance increased. When there were twenty times as many FPs as mass ROIs, the system reached a sensitivity of 89%. Adding more FP ROIs no longer improved the performance. A similar trend was observed while using only the central slice of the VOI with a maximum sensitivity of 88%.

Addition of more FP ROIs after a ratio of 25 times that of the masses again does not improve performance.

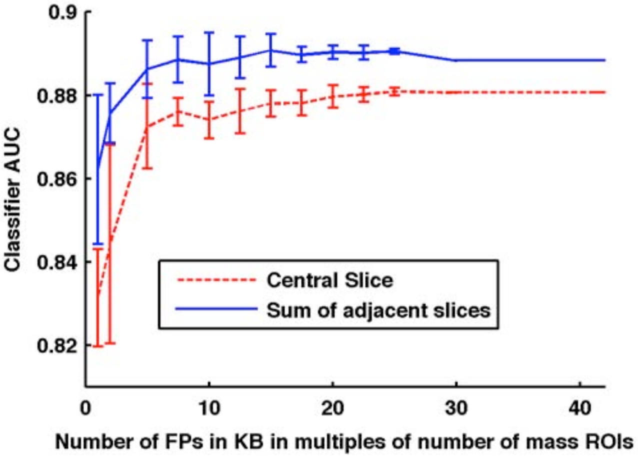


Figure 7: The figure of merit, ROC AUC is plotted as a function of increasing number of FP ROIs in the system.

Scheme B - effect of normal ROIs in the KB

Scheme B assessed the behavior of the system with the presence of normal ROIs in the KB. Figure 8 depicts this trend as a function of increasing number of normal ROIs in the system. As previously described in section, the error bars are obtained when the same data point of the graph is evaluated using 20 different subsets of the normal ROIs available. AUC increased as more normal ROIs were added to the KB and levels off at a ratio of 25 times as many normals as masses for sum of adjacent slices. The same leveling off in performance for central slice was seen with 30 times as many normals as mass ROIs. Performance was comparable to that of scheme A. Scheme B attained a maximum classifier AUC of 86% for central slice ROIs and 89% for sum of slices ROIs. As with scheme A, use of the slab ROIs did not affect performance substantially, although here in scheme B it had a more noticeable increase in performance than for scheme A.

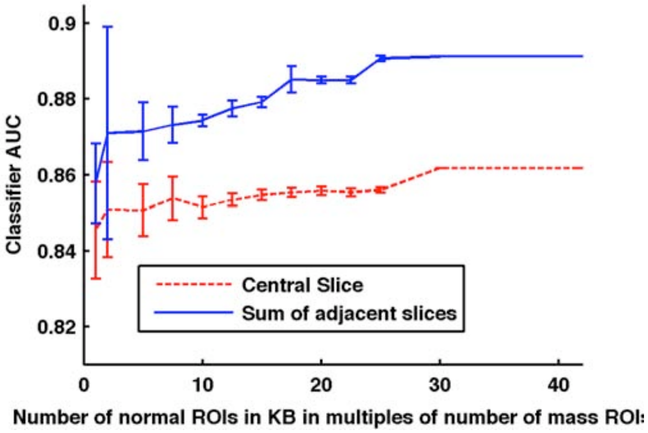


Figure 8: The figure of merit, ROC AUC is plotted as a function of increasing number of normal ROIs in the system.

As the amount of data available increases, an understanding of what constitutes an optimal KB in terms of the optimal number of FPs and/or normals will become pivotal for all practical applications. This is because similarity metrics need to be calculated for each query presented to the system with every ROI in the database. If there is nothing to be gained in terms of performance, then having more ROIs in the database simply adds to time needed for the system to generate CAD marks on a new case. To better understand the composition of such an optimal KB for tomosynthesis data, three FP reduction schemes were compared, each based on ROIs from only a single central slice versus a summed slab of slices from the first stage of the algorithm. While doing so, several trends were observed. There was no statistically significant difference in classifier performance when comparing the use of a single, central slice only versus the sum of adjacent slices, regardless of whether the AUC or partial AUC was the figure of merit. Scheme B's performance was almost the same as that of A and C, even though B doesn't use FPs in its KB. The performance of scheme B was independent of the nature of FPs generated by the first stage of the algorithm. JAFROC analyses of the system performances for the three schemes also indicate that there is no statistically significant difference between scheme B when compared against scheme A and C. Thus the results obtained for scheme B may be more robust when given either different cases or another set of unknown ROIs from these same cases that contain false positives generated by a different filter or algorithm. The performance of scheme C was between that of A and B as it added the use of FPs in its KB.

The study of the optimal balance between positive and negative cases in the KB also yielded several interesting trends. For scheme A, the system reached its maximum performance with a FP ratio of twenty times that of mass ROIs in its KB. A similar trend was observed in scheme B when the KB contained information about only masses and normal breast tissue where nearly thirty times as many normal ROIs were needed in the KB as mass ROIs. Thus it appeared that scheme B required more examples of randomly extracted normal ROIs compared to scheme A which used more suspicious normal anatomy presented in FP ROIs. Regardless of the nature of the negative, non-mass cases, both systems showed that when given increasingly larger number of non-mass ROIs in its KB, their performance increased toward an asymptote. Furthermore, we found that more non-mass ROIs than mass ROIs were needed in order for the algorithm to learn the naturally greater variability of normal breast anatomy. Both schemes displayed larger standard deviations in performance levels initially with tighter confidence levels attained as the schemes were given increasing information about the diversity of normal breast tissue.

KEY RESEARCH ACCOMPLISHMENTS

- Collected over 240 human subject data using the tomosynthesis technique.
- Optimized and applied the high sensitivity, low specificity stage of 2D CAD to projection images of the tomosynthesis data and used information theoretic similarity metrics to reduce false positives.
- Built a tomosynthesis CAD algorithm that reduces false positives in the reconstructed domain and achieves a maximal classifier sensitivity of AUC 0.89.
- Published 12 papers in peer-reviewed journals and editor-reviewed conferences.

REPORTABLE OUTCOMES
The names of the fellow (Singh) are boldfaced for emphasis.

1. **S. Singh**, G. D. Tourassi, J. A. Baker, E. Samei, J. Y. Lo, “Automated Breast Mass Detection in 3D Reconstructed Tomosynthesis Volumes: A Featureless Approach,” Accepted in Medical Physics in June 2008 (in press)

2. **S. Singh**, G. D. Tourassi, J. Y. Lo, “Effect of Similarity Metrics and ROI Sizes in Featureless Computer Aided Detection of Breast Masses in Tomosynthesis,” accepted in IWDM 2008

3. G. D. Tourassi, R. Ike, **S. Singh**, B. Harrawood, “Evaluating the Effect of Image Processing on an Information-Theoretic CAD system in Mammography,” Academic Radiology, May 2008

4. G. D. Tourassi, B. Harrawood, **S. Singh**, J. Y. Lo, Floyd CE, “Evaluation of Information-Theoretic Similarity Measures for Content Based Retrieval and Detection of Masses in Mammograms,” Medical Physics, January 2007, 140- 150, Volume 34

5. G. D. Tourassi, B. Harrawood, **S. Singh**, J. Y. Lo, “Information- Theoretic CAD System in Mammography: Entropy- Based Indexing for Computational Efficiency and Robust Performance,” Medical Physics, August 2007, 3193- 3204, Volume 34

6. **S. Singh**, G. D. Tourassi, A. S. Chawla, R. S. Saunders, E. Samei, J. Y. Lo, “Computer Aided Detection of Breast Masses in Tomosynthesis Reconstructed Volumes Using Information-Theoretic Principles,” Medical Imaging: CAD 2008

7. R. Ike, **S. Singh**, G. D. Tourassi, “Effect of ROI Size on the Performance of an Information-Theoretic CAD System in Mammography: Multisize Analysis Fusion,” Medical Imaging: CAD 2008

8. A. S. Chawla, E. Samei, R. S. Saunders, J. Y. Lo, **S. Singh**, “Optimized Acquisition Scheme for Multi-projection Correlation Imaging of Breast,” Medical Imaging: Image Perception 2008

9. J. Y. Lo, **S. Singh**, J. T. Dobbins, E. Samei, “New Developments in Digital Breast Tomosynthesis,” American Associates of Physicists in Medicine (AAPM), 2007

10. **S. Singh**, G. Tourassi, **J. Y. Lo**, “Breast Mass Detection in Tomosynthesis Projection Images Using Information Theoretic Similarity Measures.” Medical Imaging: CAD, 2007

11. A. Jerebko, Y. Quan, N. Merlet, E. Ratner, **S. Singh**, J. Y. Lo, Arun Krishnan, “Feasibility study of breast tomosynthesis CAD system.” Medical Imaging: CAD 2007

12. J. Y. Lo, J. A. Baker, J. Orman, T. Mertelmeier, **S. Singh**, “Breast Tomosynthesis: Initial Clinical Experience with 100 Human Subjects.” Radiological Society of North America (RSNA) 2006

CONCLUSIONS

A CADe system for breast tomosynthesis was developed which attained promising results over a dataset of one hundred human subjects consisting of twenty-five mass cases. A Difference of Gaussian (DoG) filter was used in the projection domain to identify initial suspicious locations. We then reconstructed these suspicious locations identified by CADe in the projection domain using a shift and add reconstruction method to get results in the 3D reconstructed domain. Centroids of the volume of interests were identified and used to extract ROIs that were subjected to a false positive reduction stage utilizing information theory principles.

The best overall system performance was achieved while using a knowledge base consisting of mass and false positive ROIs. Adding normal ROIs in addition to or in place of the false positives resulted in the same sensitivity but slightly worse specificity, but may represent more generalizable results as doing so decreased the dependence on specifics of this detection algorithm. In conclusion, this CAD system was based on a human subject data set and used an innovative false positive reduction scheme of feature-less information theory based similarity metrics, and demonstrated promising results for mass lesion detection.

REFERENCES

1. G. D. Tourassi, B. Harrawood, **S. Singh**, J. Y. Lo, Floyd CE, “Evaluation of Information-Theoretic Similarity Measures for Content Based Retrieval and Detection of Masses in Mammograms,” Medical Physics, January 2007, 140- 150, Volume 34

APPENDICES

The peer-reviewed paper from Medical Physics currently in press (reportable outcome #1) and IWDM conference proceeding (reportable outcome #2) are attached as appendices to this report.

Appendix

1. **S. Singh**, G. D. Tourassi, J. A. Baker, E. Samei, J. Y. Lo, “Automated Breast Mass Detection in 3D Reconstructed Tomosynthesis Volumes: A Featureless Approach,” Accepted in Medical Physics in June 2008 (in press)

2. **S. Singh**, G. D. Tourassi, J. Y. Lo, “Effect of Similarity Metrics and ROI Sizes in Featureless Computer Aided Detection of Breast Masses in Tomosynthesis,” accepted in IWDM 2008

3. G. D. Tourassi, R. Ike, **S. Singh**, B. Harrawood, “Evaluating the Effect of Image Processing on an Information-Theoretic CAD system in Mammography,” Academic Radiology, May 2008

4. G. D. Tourassi, B. Harrawood, **S. Singh**, J. Y. Lo, Floyd CE, “Evaluation of Information-Theoretic Similarity Measures for Content Based Retrieval and Detection of Masses in Mammograms,” Medical Physics, January 2007, 140- 150, Volume 34

5. G. D. Tourassi, B. Harrawood, **S. Singh**, J. Y. Lo, “Information- Theoretic CAD System in Mammography: Entropy- Based Indexing for Computational Efficiency and Robust Performance,” Medical Physics, August 2007, 3193- 3204, Volume 34

6. **S. Singh**, G. D. Tourassi, A. S. Chawla, R. S. Saunders, E. Samei, J. Y. Lo, “Computer Aided Detection of Breast Masses in Tomosynthesis Reconstructed Volumes Using Information-Theoretic Principles,” Medical Imaging: CAD 2008

7. R. Ike, **S. Singh**, G. D. Tourassi, “Effect of ROI Size on the Performance of an Information-Theoretic CAD System in Mammography: Multisize Analysis Fusion,” Medical Imaging: CAD 2008

8. A. S. Chawla, E. Samei, R. S. Saunders, J. Y. Lo, **S. Singh**, “Optimized Acquisition Scheme for Multi-projection Correlation Imaging of Breast,” Medical Imaging: Image Perception 2008

9. **S. Singh**, G. Tourassi, **J. Y. Lo**, “Breast Mass Detection in Tomosynthesis Projection Images Using Information Theoretic Similarity Measures.” Medical Imaging: CAD, 2007

10. A. Jerebko, Y. Quan, N. Merlet, E. Ratner, **S. Singh**, J. Y. Lo, Arun Krishnan, “Feasibility study of breast tomosynthesis CAD system.” Medical Imaging: CAD 2007

Automated Breast Mass Detection in 3D Reconstructed Tomosynthesis Volumes: A Featureless Approach

Swatee Singh[1-3], Georgia D. Tourassi[1,2,4], Jay A. Baker[2], Ehsan Samei[1-5], Joseph Y. Lo[1-4]

[1]Duke Advanced Imaging Laboratories
Department of Radiology
Duke University Medical Center
Durham, NC 27705

[2]Department of Radiology
Duke University Medical Center
Durham, NC 27710

[3]Department of Biomedical Engineering
Duke University
Durham, NC 27708

[4]Department of Medical Physics
Duke University
Durham, NC 27705

[5]Department of Physics
Duke University
Durham, NC 27710

ABSTRACT

The purpose of this study was to propose and implement a computer aided detection (CADe) tool for breast tomosynthesis. This task was accomplished in two stages – a highly sensitive mass detector followed by a false positive (FP) reduction stage. Breast tomosynthesis data from 100 human subject cases were used; of which 25 subjects had one or more mass lesions and the rest were normal. For stage 1, filter parameters were optimized via a grid search. The CADe identified suspicious locations were reconstructed to yield 3D CADe volumes of interest. The first stage yielded a maximum sensitivity of 93% with 7.7 FPs/ breast volume. Unlike traditional CADe algorithms in which the second stage FP reduction is done via feature extraction and analysis, instead information theory principles were used with mutual information as a similarity metric. Three schemes were proposed, all using leave-one-case-out cross validation sampling. The three schemes, A, B and C, differed in the composition of their knowledge base of regions of interest (ROIs). Scheme A's knowledge base was comprised of all the mass and FP ROIs generated by the first stage of the algorithm. Scheme B had a knowledge base that contained information from mass ROIs and randomly extracted normal ROIs. Scheme C had information from three sources of information – masses, FPs and normal ROIs. Also, performance was assessed as a function of the composition of the knowledge base in terms of the number of FP or normal ROIs needed by the system to reach optimal performance. The results indicated that the knowledge base needed no more than twenty times as many FPs and thirty times as many normal

ROIs as masses to attain maximal performance. The best overall system performance was 85% sensitivity with 2.4 FPs per breast volume for scheme A, 3.6 FPs per breast volume for scheme B and 3 FPs per breast volume for scheme C.

Keywords: Computer Aided Detection, Tomosynthesis, Mass Detection, Projection Images, Reconstructed Volume, Information Theory, Mutual Information, Knowledge Base, Breast Imaging, Mammography, Masses

I. INTRODUCTION

Mammography is currently the most effective early-detection tool for breast cancer screening. To provide a reliable and efficient second-reader to aid breast-imaging radiologists, recent research has been directed towards developing computer-aided detection (CADe) tools for mammography.[1-17] Although these tools have shown promise in identifying calcifications, detecting masses has proven relatively more difficult primarily due to presence of dense overlying tissue in a mammogram. Breast tomosynthesis has the potential to improve detection and characterization of breast masses by removing overlapping dense fibroglandular tissue. These systems provide 3D slice images from a modified full field digital mammography system which acquires a limited-angle cone beam CT scan under mammography positioning. Recent studies such as that by Poplack et al[18] demonstrated decreased recall rate and superior image quality for tomosynthesis versus conventional mammography. The goal of tomosynthesis to provide 3D information at comparable dose, resolution, and patient throughput to mammography, and with lower cost and hardware requirements compared to alternatives such as breast Computed Tomography or breast Magnetic Resonance Imaging. However, with tomosynthesis, instead of the traditional 4 mammography views per case, the radiologist must interpret a large volume of data per breast volume. Given this constraint, the role of CADe is especially important in breast tomosynthesis. If this modality is ever intended to replace mammography as a screening tool, then a CADe algorithm that presents the radiologist with initial cues could potentially become indispensable to maintain current clinical workflow. In fact, investigators in CT colonography have already begun to show that CADe can potentially ease radiologist workflow with large 3D datasets.[19]

Previous CADe studies have reported CADe models for breast tomosynthesis. Reiser et al[20] have modified their 2D mammography algorithms to work with 3D tomosynthesis data. Their dataset consisted of 36 cases wherein 35 were biopsy proven malignant masses and 1 was benign. The training and testing set were the same, resulting in sensitivity of 90% with 1.5 false positives (FP) per breast volume.[20] Chan et al[21] have combined information from 2D projection images with 3D volumes in 52 cases wherein 41 were malignant masses and 11 were benign. They reported sensitivities of 80% and 90% at an average FP rate of 1.2 and 2.3 per breast respectively while using a leave one out cross validation scheme. Comparable performances have been reported in other studies using smaller datasets.[22-24]

We propose to build a CADe scheme for tomosynthesis, incorporating unique preprocessing techniques and information theory methods. The CADe system in this study has two key components: 1) a highly sensitive mass detector, and 2) statistical models designed to reduce false-positives. The 'high-sensitivity, low specificity' stage of the proposed algorithm is the first component and is comprised of a Difference of Gaussians (DoG) filter. The second, 'high-sensitivity, high-specificity' stage of the algorithm is comprised of false positive (FP) reduction using information theory principles. Previous 2D algorithms for mammograms that use information theory and similarity metrics to reduce false positives have shown that the ability of the system to optimally perform such a task is dependent on the nature of the 'known' examples in the database available to it as the learning cases.[25,26] Therefore, further analysis is performed to identify the optimal knowledge base for our system. Three FP reduction schemes were evaluated that differ in the kind of information available for the task of false positive reduction. Finally, to explore if there are performance increases to be realized if more signal information was given to the system, two variants of the FP reduction system were compared – using only the central reconstructed slice of the CADe suspicious location versus using a summed slab of slices.

II. METHODS AND MATERIALS

A. Dataset

Our dataset was collected using a prototype breast tomosynthesis system Mammomat Novation TOMO[a] by Siemens Medical Solutions (Erlangen, Germany), which acquires 25 projection images over a 50-degree angular range in approximately 13 seconds. The projection images are acquired using an amorphous selenium direct digital detector with a large surface area (24x30 cm) and with an 85-micron pixel pitch. Projection images of 2816x3584 pixels with 2x1 pixel binning in the tube motion direction are acquired by this system at the rate of 2 images/second. Institutional review board approval was obtained, and informed consent was required and obtained for all subjects. This study was compliant with the Health Insurance Portability and Accountability Act. The protocol called for bilateral MLO views to be acquired in screening cases, while bilateral MLO and CC views were acquired for diagnostic and biopsy cases. An MQSA dedicated breast radiologist with over fifteen years of experience interpreted these cases in blinded readings. The gold standard was established from information available from all modalities for a subject including mammography and, when available, ultrasound and MRI for non-biopsied lesions, while biopsied lesions resulted in definitive histopathologic truth. One hundred human subject cases were used wherein there were twenty-five mass cases and seventy-five normal cases. All of these subjects were recruited at Duke University Medical Center in Durham, NC and had an average age of 57

[a] Caution: Investigational Device. Limited by US Federal law to investigational use. The information about this product is preliminary. The product is under development and is not commercially available in the US; and its future availability cannot be ensured.

years. Approximately 24% of the subjects had breast density of 25%, 20% were 50% dense, 46% were 75% dense and 10% were considered to have 100% dense breast. 83% of these subjects were Caucasian, 13% African-American and 4% identified themselves as either Hispanic or Asian. Due to some unilateral cases, a total of 192 scans were evaluated. The twenty-five mass cases contained twenty-eight lesions of which ten were biopsy-proven malignant lesions and the rest were benign. Focal asymmetries and calcifications were excluded from this study. Average lesion size is approximately (100x100) ± 41 pixels or (8.5 x 8.5) ± 3.5 mms.

B. Overview of the CADe system

The CADe scheme is comprised of two distinct stages – the 'high-sensitivity, low specificity' stage wherein ROIs are extracted followed by the 'high-sensitivity, high-specificity' stage that uses information theory principles to reduce false positives. We worked with the raw projection images with only standard detector preprocessing including dead pixel and uniformity correction. A schematic of the 2 stages can be seen in Figures 1 and 3. The system performs the following steps:

a. For each of the 25 projection images, the breast edge was detected by estimating an optimal threshold to distinguish the class distributions of the foreground and background pixels. Only information inside the breast boundary was preserved and was subsequently filtered.
b. Threshold segmented, filtered projection images from step (a) to yield CADe suspicious locations in 2D.

c. Reconstruct only the CADe suspicious locations generated by step (b) via shift and add reconstruction method to yield 3D volumes of CADe suspicious locations.[27]

d. Locate the center of the CADe reconstructed suspicious locations in 3D and map to the filtered backprojection (FBP) 3D reconstructed volume used during radiologist interpretation.

e. Extract ROIs from FBP reconstructed slices at the locations specified in step (d).

f. Implement various FP reduction schemes to attain final system performances.

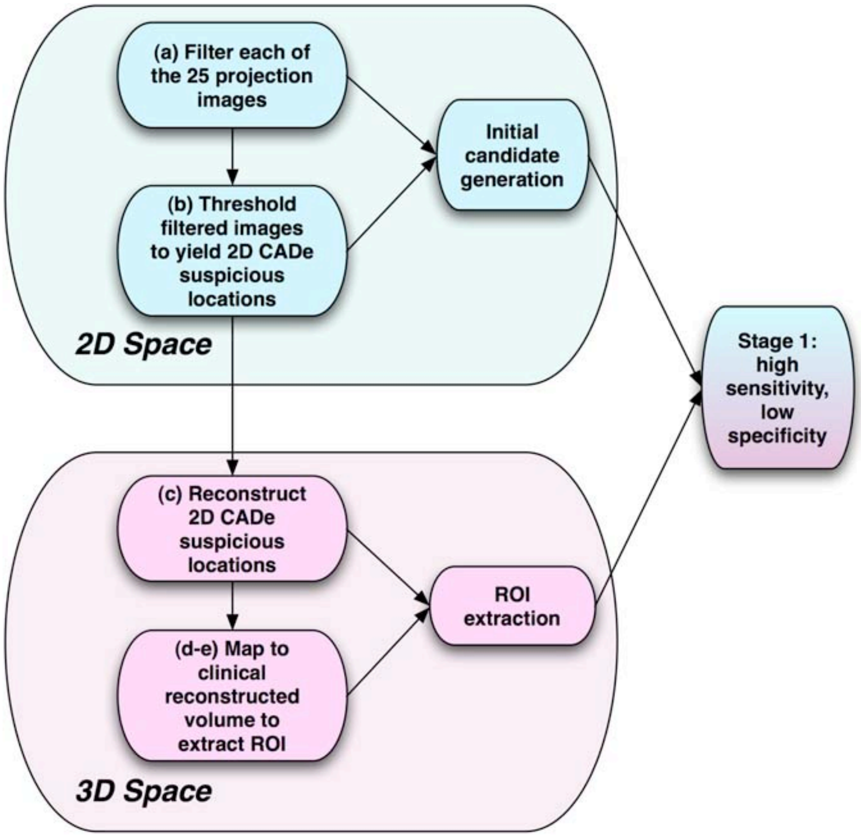

(a) Filter each of the 25 projection images
Initial candidate generation
(b) Threshold filtered images to yield 2D CADe suspicious locations
2D Space
Stage 1: high sensitivity, low specificity
(c) Reconstruct 2D CADe suspicious locations
ROI extraction
(d-e) Map to clinical reconstructed volume to extract ROI
3D Space

Figure 1: Stage 1 – the filtration and ROI extraction or the 'high-sensitivity, low specificity' stage of the CADe algorithm

1. Stage 1 – Filtration and ROI extraction

For each breast view, the 25 projection images were filtered using a Difference of Gaussians (DoG) filter.[28-30] The DoG filter in two dimensions is achieved by subtracting a rotationally symmetric, two-dimensional Gaussian with width parameter σ_1 from another rotationally symmetric, two-dimensional Gaussian with width parameter σ_2. Mathematically, the filter template *w* is defined as:

$$w = G_1(r \mid \sigma_1) - G_2(r \mid \sigma_2) \tag{1}$$

where

$$G_i(r \mid \sigma_i) = \frac{1}{\left(\sigma_i \sqrt{2\pi}\right)} e\left\{-\left(\frac{r^2}{\sigma_i^2}\right)\right\} \tag{2}$$

where *r* is the distance to the origin and σ_i is the constituent width parameter of the filter template. Of note here is the relationship between the two standard deviations where $\sigma_1 < \sigma_2$.

Each of the filtered projection images was then subjected to adaptive thresholding to yield CADe suspicious locations in each of the projections. In that process, the thresholds for each of the projection images were dynamically selected by starting with the top 10% of the pixel values of the filtered projection image resulting in an initial set of CADe suspicious locations. Further drops in the threshold resulted in either an increase in the area of the initial suspicious locations or in the formation of new ones. The threshold was thus dropped as low as possible without merging together any two suspicious locations. For dense breasts, this threshold often included approximately 15% of the top pixel values, while for fatty breasts the thresholds were generally selected at about 25% of the top pixel values. Only the segmented 2D projection images thus obtained were shifted and added using the acquisition angle and known geometry to yield 3D locations of the volume of interest (VOI) of just the CADe locations.[31]

A 3x3x3 connectivity rule was used to yield CADe suspicious locations in 3D space making it possible to determine location and shape of the object of interest. Specifically, every pixel in each of the slices of the reconstructed slices of the CADe suspicious locations was assigned to a VOI using its proximity to a cluster of pixels. This resulted in a set of VOIs for every scanned breast view. Since the shift and add reconstruction algorithm did not have any measures in its implementation to prevent out of plane blur, the resulting reconstructed CADe suspicious volumes from the first stage had significant blur in planes other than where the centroid of the volume of interest lies, resulting in a starburst shape wherein the true object lies in the plane where the contributions from all the projection images come into focus. Thus, it is assumed that a mass came into focus in the plane with the least cross-sectional area of the volume obtained after reconstruction. False positives due to overlapping tissue in just a few projection images should result in weaker 3D reinforcement of signals. An example of such a reconstruction is shown in Figure 2. This 3D location of the volume of interest was then compared to the radiologist-determined ground truth to determine if a given CADe location is a true positive or a false positive. To determine whether a CADe suspicious location is indeed a TP, the following rule was used:

$$\text{If}\left(\frac{A(CADe)\cap A(Truth)}{A(CADe)\cup A(Truth)}\right) > 0.3,\ \text{then TP}$$

where A(CADe) is the area of the CADe location, and A(Truth) is the area of the truth location.

The optimization of the first high-sensitivity, low-specificity stage of the algorithm was done using all available cases, as there were not enough mass cases in our database to establish separate reasonably sized testing and training sets. The figure of merit was the maximum sensitivities as a function of the 2 DoG parameters, σ_1 and σ_2. For the lesions in our database, the average size is approximately 100x100 pixels (8.5 x 8.5 mm). A search was therefore performed wherein the filter parameters were varied from 32 to 152 pixels (2.7 to 12.92 mm) to bracket that size.

Once the algorithm had identified initial candidates for mass detection by giving the X, Y and Z location of the centroid of the volume of interest, regions of interests (ROIs) were extracted from the reconstructed breast slice images obtained by filtered backprojection (FBP) which yielded 1 mm thick slices with 85x85 micron pixel pitch.[32,33] Also shown in the Figure 2 is the corresponding lesion ROI that was extracted from the FBP reconstructed volume. The FP reduction scheme therefore was based upon the same reconstructed image data as used by radiologists. Two sets of ROIs were extracted to assess the effect of information from one versus many slices. In the first set, 256x256 pixel ROIs (22 x 22 mm) centered at the central slice containing the suspicious CADe location were extracted. For the second set, 256x256 ROIs representing the summed slab of 5 slices (5 mm) were extracted. Since lesions typically span multiple reconstructed slices, these two sets investigated whether giving more ‘signal’ to the false positive reduction scheme resulted in an improvement in performance. Use of a slab would also reduce the impact of slight errors of localization in the Z direction.

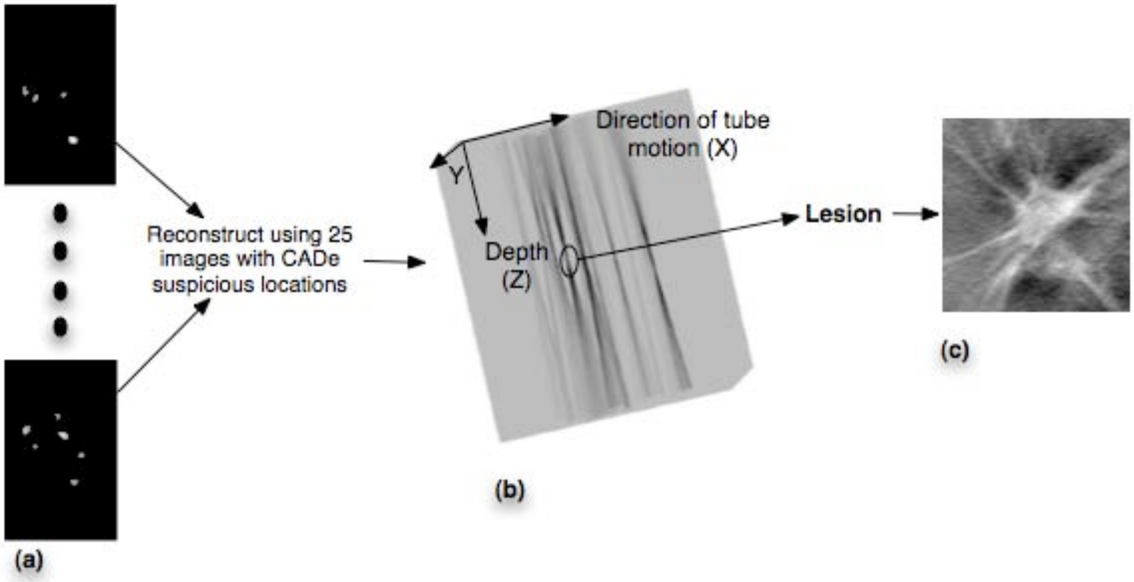

Reconstruct using 25 images with CADe suspicious locations
Direction of tube motion (X)
Y
Depth (Z)
Lesion
(a)
(b)
(c)

Figure 2: (a) 25 CADe suspicious locations in 2D for subject 33 (b) Reconstructed CADe suspicious locations using the images in ‘a.’ Significant out of plane blur is observed in Z direction. (c) A 256x256 ROI centered at the X, Y location at the depth with the sharpest focus is extracted from the FBP reconstructed volumes and shown in ‘b.’

2. Stage 2 – False positive reduction

Information theory principles were used to reduce false positives (FPs) in the second stage of the algorithm. The fundamental quantities of information theory are entropy and relative entropy. For any probability distribution, entropy is defined as a quantity that follows an intuitive notion of a measure of information. In other words, entropy, among other measures such as variance etc, is a way to quantify the uncertainty involved in a random variable. This notion is extended to define ‘mutual information’ which is a measure of the amount of information one random variable contains about another. Hence, mutual information is a reduction in the uncertainty of one random variable due to the knowledge gleaned from observing the other random variable. Mathematically, it is given by the following relation[34]:

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{3}$$

where *X* and *Y* are two random variables, $p(x,y)$ is their joint probability mass function because this is a discrete rather than continuous random variable and $p(x)$ and $p(y)$ are the marginal probability mass functions of *X* and *Y*.

Traditionally CADe schemes measure, among others, morphological and texture features of a suspicious location for subsequent false positive reduction using trainable classifiers. Research has been done by Suzuki et al[35-38] towards alternative approaches to FP reduction by using massive training artificial neural networks. This study used mutual information as a similarity metric for false positive reduction that relies completely on the statistical properties of the image histograms and the relationship between pixels of an image. Furthermore, information theoretic similarity measures make no assumptions about the underlying image distributions, which may be advantageous given the relatively small number of lesions in our dataset. The theoretical approach adopted in this study has been presented previously[39-41] for 2D mammograms. This study extended the concept for 3D reconstructed slices and slabs.

An information theory based system compares an unknown query ROI to every ROI in its ‘knowledge base’ (KB) using a similarity metric such as mutual information. Similarity metrics are then combined using a decision index[41] given in equation 4.

$$D(Q) = \frac{1}{m}\sum_{j=1}^{m} MI(Q,M_j) - \frac{1}{n}\sum_{j=1}^{n} MI(Q,N_j) \tag{4}$$

where *Q* is the query ROI, $MI(\bullet,\bullet)$ is the mutual information between the query *Q* and the ROI in the KB. M_j and N_j are the j^{th} mass and normal ROI respectively in a KB that contains a total of *m* mass and *n* normal ROIs. By applying various thresholds on these indices for all cases in the database the performance can be studied as a receiver operating characteristic (ROC) curve. Both area under curve (AUC) and partial area under curve (pAUC) above 90% sensitivity were measured nonparametrically.[42,43] To estimate the two-sided p-value for the central slice versus sum of adjacent slices datasets for each scheme, a set of cases was bootstrapped to estimate the difference in performance. This was repeated to obtain an estimate of the difference distribution.

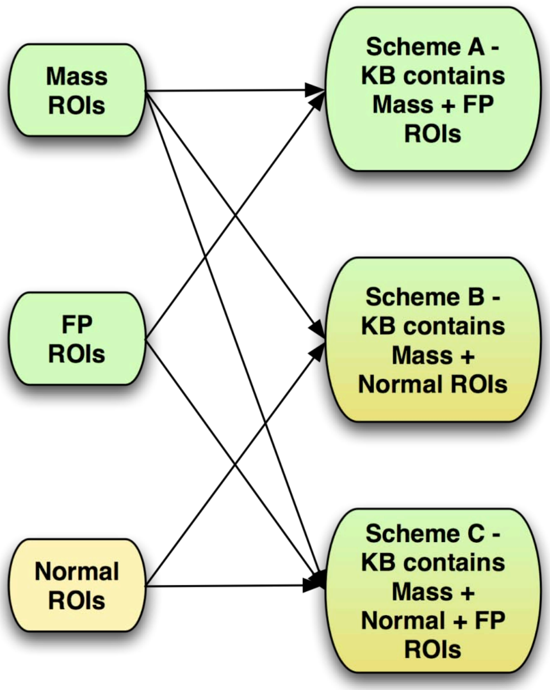

Mass ROIs
FP ROIs
Normal ROIs
Scheme A - KB contains Mass + FP ROIs
Scheme B - KB contains Mass + Normal ROIs
Scheme C - KB contains Mass + Normal + FP ROIs

Figure 3: Composition of Knowledge Base of false positive reduction stage of the CADe algorithm (a) scheme 'A' KB composition (b) schemes 'A' and 'B' composition

The performance of an information theory based system is dependent on the composition of the knowledge base. The first stage of the algorithm generates ROIs that are either mass lesions or FPs. Of note here is the imbalance in the number of lesion ROIs when compared to the total number of FPs generated by the first stage. Given this imbalance, it is imperative to explore the effect of knowledge about normal breast parenchyma represented by those FPs. This was studied in two ways. First, an increasing number of FPs was sampled from all FPs available while holding the number of true positives constant, thus decreasing the ratio of mass ROIs in the KB and progressively giving the system more indirect 'knowledge' of normal breast parenchyma. The second approach is to provide the system with direct information about normal breast parenchyma via randomly selected normal ROIs instead of suspicious FP regions generated by a CADe algorithm. Since these ROIs were extracted from random locations from within the breast volume there is a potential for some overlap with FPs generated by the first stage of the algorithm. Varying the number of mass ROIs in the knowledge base can also change composition of the knowledge base. However, given that our database consists of a limited number of mass ROIs, its effect was not studied in this experiment.

Three schemes were therefore developed to investigate the optimal ratio of normal and false positive ROIs in the knowledge base, as shown in Figure 3. In scheme A, FP reduction was done using a KB containing ROIs from the CADe algorithm's first stage. These ROIs were either mass ROIs or FPs. In scheme B, the KB contained only mass ROIs and randomly selected normal ROIs from well-separated depths in all the normal cases' reconstructed volumes. A total of 1390 such normal ROIs were extracted for this study. To access performance of the scheme A classifier, a leave-one-case-out validation scheme was used. Thus, for every ROI that was presented to the system as a query ROI of unknown pathology, all other ROIs generated from that specific subject's reconstructed volumes were excluded from the KB. For scheme B, all the FPs of the first stage of the algorithm served as queries to the system to assess its specificity. Sensitivity for scheme B was evaluated using a leave-one-case-out sampling scheme on all available ROIs that contained a mass. Thus the system has no knowledge of FP ROIs in its KB and hence the performance is not dependent on the nature of FP lesions generated by the first stage of the algorithm. Finally, scheme C included information from all three sources, (1) masses (2) CADe generated FPs (3) normal breast tissue, combined into a single KB. Analysis was done in a leave-one-case-out manner for this KB as well. In the end, the scores for all ROIs thus obtained from various schemes were then combined using the decision index given by equation 4.

III. RESULTS

A. Optimization of stage 1

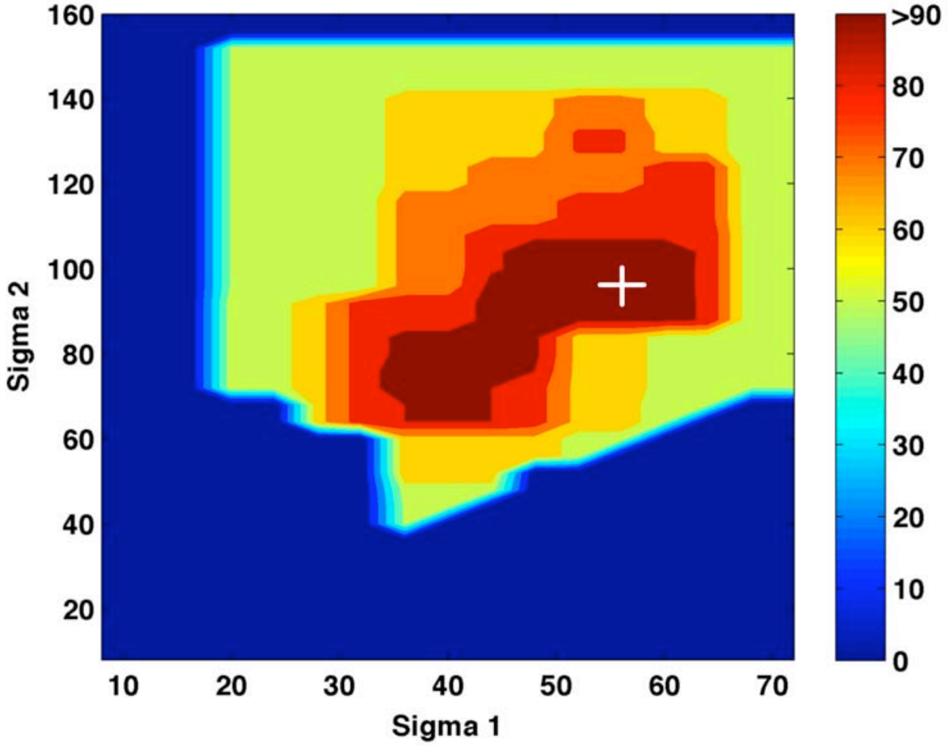


Figure 4: Sensitivity as a function of the 2 filter parameters for stage 1 of the algorithm. The combination marked by the '+' was chosen, yielding 93% sensitivity with 7.7 FPs/ breast volume.

Optimization of the first high-sensitivity, low-specificity stage of the algorithm involved a grid search over the 2 DoG parameters, σ_1 and σ_2. Maximum sensitivity for each combination is shown in Figure 4. The parameter sets that were not explored are represented with a zero percent sensitivity. While the FP rate for each parameter set was recorded, no specific optimization for the FP rate was performed. There were 2 distinct areas with high reported sensitivities, centered at σ_1 and σ_2 pairs of 40/ 72 (3.4/ 6.12 mm) and 56/ 96 pixels (4.76/ 8.16 mm) with 9.3 and 7.7 FPs/ breast volume respectively. The parameters 56/ 96 yielded fewer false positives and were therefore picked for further analysis of stage 2. Thus, the first stage of the algorithm yielded a maximum sensitivity of 93% and 1472 FPs resulting in a FP rate of 7.7 FPs per breast volume. All available cases were used for the optimization of this stage due to the small size of the dataset resulting in the possibility of a positive bias in the reported performance of the proposed algorithm.

B. Optimization of the FP reduction stage

B. 1. Scheme A - effect of FP ROIs in the KB

Scheme A seeks to differentiate between a mass and a FP query. A plot of the ROC AUC as a function of increasing number of FPs is presented in Figure 5, where the x-axis shows number of FPs as multiples of the total number of mass ROIs while using the scheme A classifier. The error bars are obtained by simple random sampling[25,44] from all the available FPs of the first stage. 20 subsets of the FP ROIs were generated for each data point on the graph. Each subset was selected without replacement after randomization between subsets. When the sum of adjacent slices were used, as the number of FPs was increased the performance increased. When there were twenty times as many FPs as mass ROIs, the system reached a sensitivity of 89%. Adding more FP ROIs no longer improved the performance. A similar trend was observed while using only the central slice of the VOI with a maximum sensitivity of 88%. Addition of more FP ROIs after a ratio of 25 times that of the masses again does not improve performance. It should be noted that as the number of multiples of

FP in the KB increases, the error bars in Figure 5 will also decrease because of increasing overlap in selected FP ROIs for each draw.

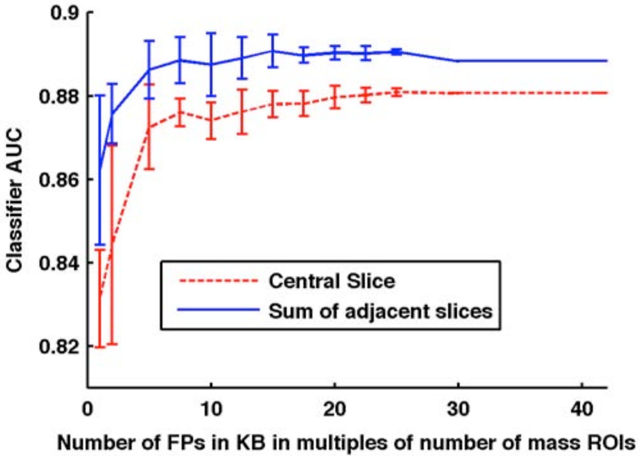


Figure 5: The figure of merit, ROC AUC is plotted as a function of increasing number of FP ROIs in the system.

B. 2. Scheme B - effect of normal ROIs in the KB

Scheme B assessed the behavior of the system with the presence of normal ROIs in the KB. Figure 6 depicts this trend as a function of increasing number of normal ROIs in the system. As previously described in section B.1., the error bars are obtained when the same data point of the graph is evaluated using 20 different subsets of the normal ROIs available. AUC increased as more normal ROIs were added to the KB and levels off at a ratio of 25 times as many normals as masses for sum of adjacent slices. The same leveling off in performance for central slice was seen with 30 times as many normals as mass ROIs. Performance was comparable to that of scheme A. Scheme B attained a maximum classifier AUC of 86% for central slice ROIs and 89% for sum of slices ROIs. As with scheme A, use of the slab ROIs did not affect performance substantially, although here in scheme B it had a more noticeable increase in performance than for scheme A.

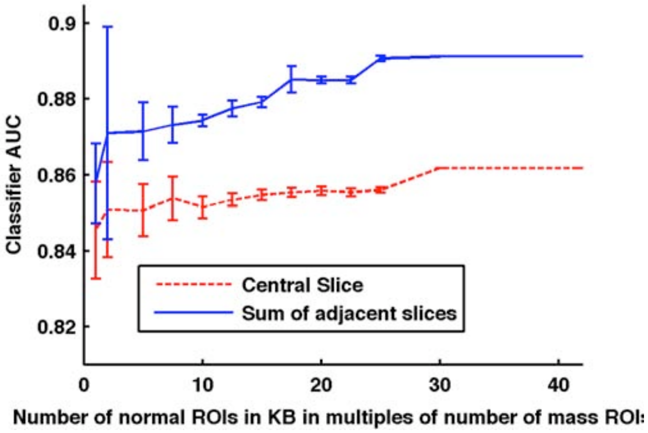


Figure 6: The figure of merit, ROC AUC is plotted as a function of increasing number of normal ROIs in the system.

C. Classifier Performances

Table 1 presents overall classifier performance for all schemes. As implemented, summing adjacent slices did not improve the classifier performance in a statistically significant way compared to using only the single, central slice ROI for any of the schemes evaluated, either for AUC or partial AUC. Shown in Figure 7 are the ROCs and partial ROCs of just the central slice classifiers of all schemes.

Table 1: Classifier performance for a KB containing mass and FP ROIs (scheme A), mass and normal ROIs (scheme B) and when the KB contains ROIs from all 3 sources – mass, FP and normal ROIs (scheme C). The AUC and pAUC for both the central slice and the sum of adjacent slices and their corresponding p-values for all schemes is shown.

Scheme	Central Slice only		Sum of Adjacent slices		p-value	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
A	0.88 +/- 0.02	0.49 +/- 0.09	0.89 +/- 0.03	0.46 +/- 0.10	0.3	0.2
B	0.86 +/- 0.03	0.41 +/- 0.09	0.89 +/- 0.03	0.36 +/- 0.10	0.5	0.2
C	0.87 +/- 0.02	0.45 +/- 0.09	0.88 +/- 0.03	0.41 +/- 0.10	0.43	0.19

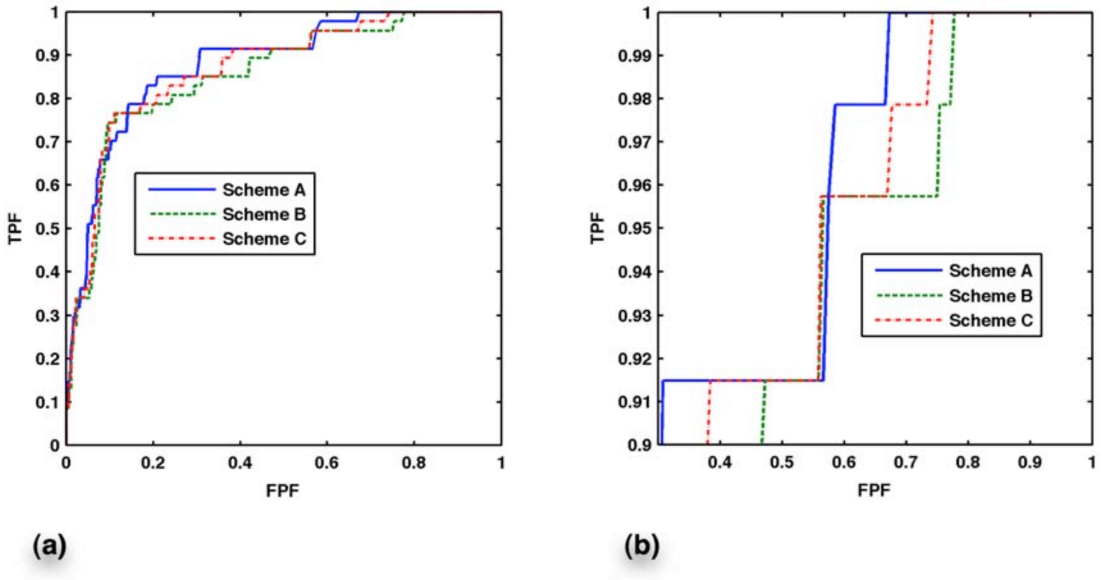


Figure 7: (a) Non-parametric ROC curves of the central slice classifier for schemes A, B, and C (b) Partial ROC curves for sensitivity greater than 0.9 for the three schemes

Sensitivity when plotted as a function of the average FP rate while the decision threshold is varied results in the Free-Response Receiver Operating Characteristic (FROC) curve. Figure 8 shows the system FROCs prior to FP reduction as well as after FP reduction for schemes A, B and C. These were obtained by varying the decision threshold over classifier outputs of the central slice classifiers of the three schemes starting with a threshold set at 91.5% sensitivity. For each scheme, the threshold was then progressively dropped to obtain the entire curve. Scheme A outperformed others in terms of FPs per breast volume at equivalent sensitivity. At an operating point of 91.5%, scheme A was successfully able to discard 69% of the FPs per breast volume, scheme B correctly eliminated 53% of the FPs per breast volume, and lastly, scheme C was able to correctly discard 62% of the FPs per breast volume. The final performances were a sensitivity of 85% at 2.4 FPs per breast volume, 3.6 FPs per breast volume, and 3 FPs per breast volume for schemes A, B and C respectively. The Jackknife Free-Response Receiver Operating Characteristic (JAFROC)[45] was used to evaluate these FROC curves. None of the differences between the FROC curves of the three schemes studied were statistically significant. A human subject example from subject 122 is shown in Figure 9. While this subject had 5 FPs in total only 2 reconstructed slices containing 1 TP and 2 FPs are shown for illustration purposes. These results were obtained when the CADe algorithm with a scheme A central slice classifier is used while operating at 91.5% sensitivity. After FP reduction, the FP in slice 40 was eliminated, however one FP along with the TP survived in slice 36. This subject had biopsy confirmed cancer.

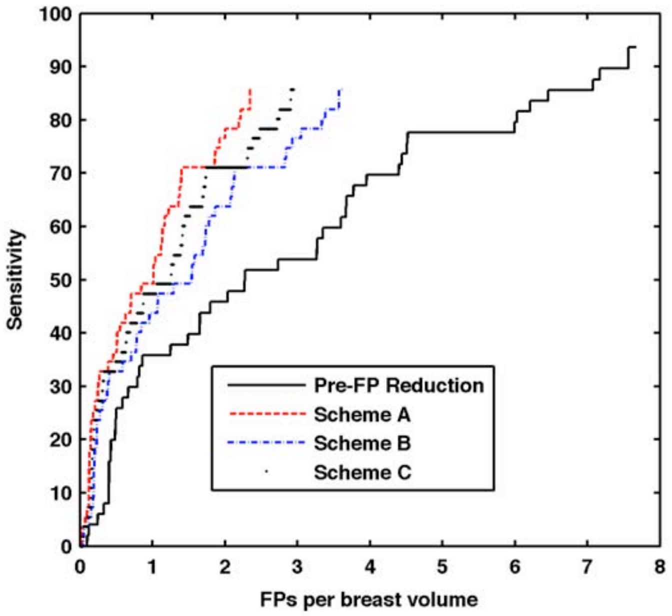


Figure 8: System FROCs. Prior to FP reduction, the system performance was at 93% sensitivity with 7.7 FPs per breast volume. Final system performances for the three schemes are depicted for the central slice classifiers.

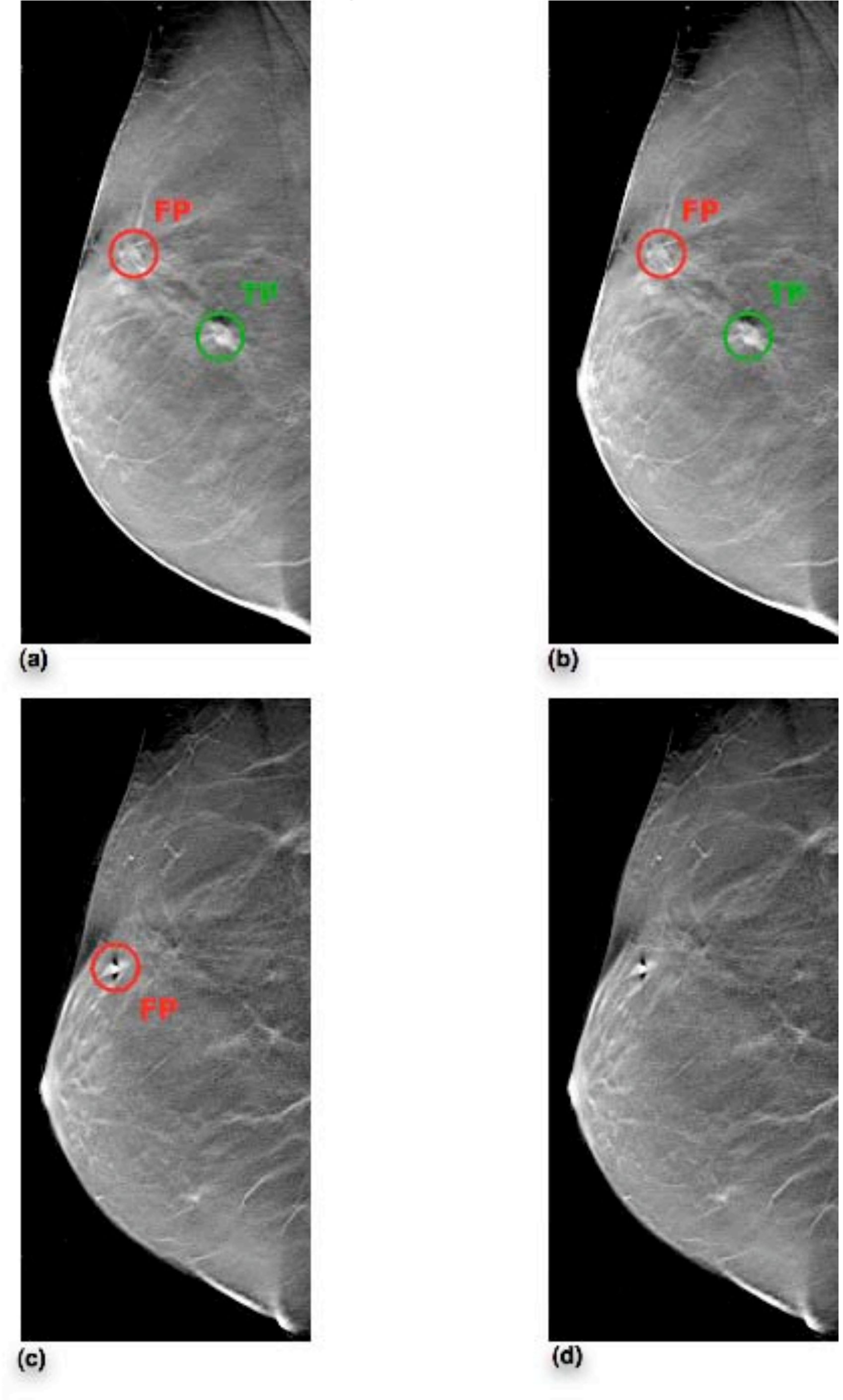


Figure 9: (a) Slice 41 prior to FP reduction (b) Slice 41 after FP reduction (c) Slice 21 prior to FP reduction (d) Slice 21 after FP reduction

Subject 122 had biopsy confirmed carcinoma. While this subject had 6 FPs in total from stage 1 of the CADe algorithm, only reconstructed slices 41 and 21 are shown in this figure for illustration. After setting the threshold for scheme A central slice classifier to operate at 91.5% sensitivity, we are able to eliminate the FP in slice 21. However, the FP in slice 41 survives along with the TP.

IV. DISCUSSION

Several CADe algorithms exist for breast tomosynthesis data in current literature. All published tomosynthesis CADe algorithms used some form of feature extraction scheme for the FP reduction stage. This study was unique in that it utilized information theory principles for this task. Given this relatively small dataset, the model still provided generalizable results when using scheme B. The generalizability here refers only to the fact that scheme B performance is independent of the nature of FPs generated by the first stage of our specific algorithm. The performance of schemes A and C can be influenced by the nature of FPs generated by other filters or another first stage of a CADe algorithm, whereas that of scheme B is independent of the kind of FPs. Additional information about mass cases would merely enhance system performance over datasets that include subjects from other geographical locations, patient populations etc. This is because inclusion of more mass cases will help the system obtain more accurate ‘knowledge’ of the various kinds of mass cases. A larger, more varied KB will have at least representative examples from all the major lesion types and will better capture the variations of various lesions.

As the amount of data available increases, an understanding of what constitutes an optimal KB in terms of the optimal number of FPs and/or normals will become pivotal for all practical applications. This is because similarity metrics need to be calculated for each query presented to the system with every ROI in the database. If there is nothing to be gained in terms of performance, then having more ROIs in the database simply adds to time needed for the system to generate CADe marks on a new case. To better understand the composition of such an optimal KB for tomosynthesis data, three FP reduction schemes were compared, each based on ROIs from only a single central slice versus a summed slab of slices from the first stage of the algorithm. While doing so, several trends were observed. There was no statistically significant difference in classifier performance when comparing the use of a single, central slice only versus the sum of adjacent slices, regardless of whether the AUC or partial AUC was the figure of merit. Scheme B's performance was almost the same as that of A and C, even though B doesn't use FPs in its KB. The performance of scheme B was independent of the nature of FPs generated by the first stage of the algorithm. JAFROC analyses of the system performances for the three schemes also indicate that there is no statistically significant difference between scheme B when compared against scheme A and C. Thus the results obtained for scheme B may be more robust when given either different cases or another set of unknown ROIs from these same cases that contain false positives generated by a different filter or algorithm. The performance of scheme C was between that of A and B as it added the use of FPs in its KB.

The study of the optimal balance between positive and negative cases in the KB also yielded several interesting trends. For scheme A, the system reached its maximum performance with a FP ratio of twenty times that of mass ROIs in its KB. A similar trend was observed in scheme B when the KB contained information about only masses and normal breast tissue where nearly thirty times as many normal ROIs were needed in the KB as mass ROIs. Thus it appeared that scheme B required more examples of randomly extracted normal ROIs compared to scheme A which used more suspicious normal anatomy presented in FP ROIs. Regardless of the nature of the negative, non-mass cases, both systems showed that when given increasingly larger number of non-mass ROIs in its KB, their performance increased toward an asymptote. Furthermore, we found that more non-mass ROIs than mass ROIs were needed in order for the algorithm to learn the naturally greater variability of normal breast anatomy. Both schemes displayed larger standard deviations in performance levels initially with tighter confidence levels attained as the schemes were given increasing information about the diversity of normal breast tissue.

Estimates of the least number of FPs or normal ROIs needed to obtain maximal performance can potentially change when more mass ROIs are added to the KB. However, a study of what that optimal number is with the current size of the dataset has lead us to the understanding that fewer FPs and normal ROIs in the KB result in greater variability in performance, and that there indeed exists a minimal ratio of these ROIs to the number of mass ROIs in the KB to attain maximal performance. Therefore, while such a ratio is likely to change with additional mass cases, there are 2 important conclusions to be drawn from these results.

These results of experiments to study optimal knowledge base composition show that for the current CADe system it is possible to attain maximal performance with little over half the number of ROIs in virtually all the three schemes. This is significant as it implies an appreciable improvement in the computational efficiency of the algorithm. The total processing time for the second stage of this algorithm that uses a LOO CV scheme is $N^2/2$. A reduction in the number of ROIs in the knowledge base by half would imply an improvement of a factor of 4 in overall computational efficiency. However, in a clinical setting the computational efficiency needs to be looked at from the point of view of a single breast volume being examined. The first stage of the algorithm generates approximately 8 CADe marks per breast view. When using a Linux Intel 2.6 GHz dual-core dual-processor system, it takes about 1 second for the system to come up with an average MI score for a single query ROI when compared against our entire knowledge base. This implies a processing time of about 2 seconds to come up with each of the two terms for equation 4 for every CADe mark from the 1st stage, and hence about 16 seconds to process the entire breast volume with 8 such potential locations generated by the 1st stage of the algorithm. Reduction in the knowledge base of half would imply a computational reduction of half, i.e. 8 seconds, in a clinical setting.

There were limitations to this study. More cases with lesions should be added to capture the diversity of breast masses. Because of the relatively small size of available dataset, the optimization of the initial filtering stage was done using all available cases with some resulting possibility of bias; addition of new cases could potentially imply a different optimal filter parameter set. The decision to sum five adjacent slices was based on the observation that most lesions spanned a space of at least 5 mm. Improvements in performance due to variation of this parameter in the algorithm has not been investigated in this study. Lastly, studies remain to be done in improving system performance by studying other similarity metrics and ROI sizes.

V. CONCLUSION

A CADe system for breast tomosynthesis was developed which attained promising results over a dataset of one hundred human subjects consisting of twenty-five mass cases. The best overall system performance was achieved while using a knowledge base consisting of mass and false positive ROIs. Adding normal ROIs in addition to or in place of the false positives resulted in the same sensitivity but slightly worse specificity, but may represent more generalizable results as doing so decreased the dependence on specifics of this detection algorithm. In conclusion, this CADe system was based on a human subject data set and used an innovative false positive reduction scheme of feature-less information theory based similarity metrics, and demonstrated promising results for mass lesion detection.

ACKNOWLEDGMENTS

This work has been supported in part by NIH/NCI R01 CA112437 and R01 CA101911, US Army Breast Cancer Research Program W81XWH-05-1-0293, and a research agreement with Siemens Medical Solutions. We would like to thank Thomas Mertelmeier PhD and Jasmina Ludwig of Siemens Medical Solutions for development of the FBP reconstruction software and helpful suggestions. Also, the authors thank the radiologists of the division of breast imaging of Duke University Medical Center for interpreting the tomosynthesis reconstructed volumes. The authors would like to extend special thanks to Brian Harrawood for his support in scientific programming for this study.

REFERENCES

[1] Y.-T. Wu, J. Wei, L.M. Hadjiiski, B. Sahiner, C. Zhou, J. Ge, J. Shi, Y. Zhang, and H.-P. Chan, "Bilateral analysis based false positive reduction for computer-aided mass detection." Med. Phys. **34**, 3334-3344 (2007).

[2] J. Wei, H.-P. Chan, B. Sahiner, L.M. Hadjiiski, M.A. Helvie, M.A. Roubidoux, C. Zhou, and J. Ge, "Dual system approach to computer-aided detection of breast masses on mammograms." Med. Phys. **33**, 4157-4168 (2006).

[3] B. Sahiner, H.-P. Chan, L.M. Hadjiiski, M.A. Helvie, C. Paramagul, J. Ge, J. Wei, and C. Zhou, "Joint two-view information for computerized detection of microcalcifications on mammograms." Med. Phys. **33**, 2574-2585 (2006).

[4] J. Wei, B. Sahiner, L.M. Hadjiiski, H.-P. Chan, N. Petrick, M.A. Helvie, M.A. Roubidoux, J. Ge, and C. Zhou, "Computer-aided detection of breast masses on full field digital mammograms." Med. Phys. **32**, 2827-2838 (2005).

[5] S. Paquerault, N. Petrick, H.P. Chan, B. Sahiner, and M.A. Helvie, "Improvement of computerized mass detection on mammograms: Fusion of two-view information." Med. Phys. **29**, 238-247 (2002).

[6] W. Qian, L. Li, and L.P. Clarke, "Image feature extraction for mass detection in digital mammography: Influence of wavelet analysis." Med. Phys. **26**, 402-408 (1999).

[7] S. Yu and L. Guan, "A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films." IEEE Trans. Med. Imaging **19**, 115-126 (2000).

[8] F. Schmidt, E. Sorantin, C. Szepesvari, E. Graif, M. Becker, H. Mayer, and K. Hartwagner, "An automatic method for the identification and interpretation of clustered microcalcifications in mammograms." Phys. Med. Biol. **44**, 1231-1243 (1999).

[9] Z. Huo, M.L. Giger, C.J. Vyborny, D.E. Wolverton, and C.E. Metz, "Computerized classification of benign and malignant masses on digitized mammograms: a study of robustness." Acad. Radiol. **7**, 1077-1084 (2000).

[10] D.M. Catarious, Jr., A.H. Baydush, and C.E. Floyd, Jr., "Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system." Med. Phys. **31**, 1512-1520 (2004).

[11] B. Zheng, Y.H. Chang, X.H. Wang, W.F. Good, and D. Gur, "Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm." Acad. Radiol. **6**, 327-332 (1999).

[12] H.P. Chan, B. Sahiner, K.L. Lam, N. Petrick, M.A. Helvie, M.M. Goodsitt, and D.D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces." Med. Phys. **25**, 2007-2019 (1998).

[13] M.A. Gavrielides, J.Y. Lo, R. Vargas-Voracek, and C.E. Floyd, Jr, "Segmentation of suspicious clustered microcalcifications in mammograms." Med. Phys. **27**, 13-22 (2000).

[14] J. Ge, B. Sahiner, L.M. Hadjiiski, H.-P. Chan, J. Wei, M.A. Helvie, and C. Zhou, "Computer aided detection of clusters of microcalcifications on full field digital mammograms." Med. Phys. **33**, 2975-2988 (2006).

[15] L. Li, Y. Zheng, L. Zheng, and R.A. Clark, "False-positive reduction in CAD mass detection using a competitive classification strategy." Med. Phys. **28**, 250-258 (2001).

[16] B. Sahiner, H.-P. Chan, N. Petrick, M.A. Helvie, and L.M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features." Med. Phys. **28**, 1455-1465 (2001).

[17] S. Singh, A.H. Baydush, B. Harrawood, and J.Y. Lo, "Mass detection in mammographic ROIs using Watson filters," SPIE Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment, San Diego, CA, **6146** (2006).

[18] S.P. Poplack, T.D. Tosteson, C.A. Kogel, and H.M. Nagy, "Digital Breast Tomosynthesis: Initial Experience in 98 Women with Abnormal Digital Screening Mammography." Am. J. Roentgenol. **189**, 616-623 (2007).

[19] H.D. Abraham and Y. Hiro, "Virtual colonoscopy: past, present,and future." Radiol. Clin. North Am. **41**, 377-393 (2003).

[20] I. Reiser, R.M. Nishikawa, M.L. Giger, T. Wu, E.A. Rafferty, R. Moore, and D.B. Kopans, "Computerized mass detection for digital breast tomosynthesis directly from the projection images." Med. Phys. **33**, 482-491 (2006).

[21] H.-P. Chan, J. Wei, Y. Zhang, R.H. Moore, D.B. Kopans, L. Hadjiiski, B. Sahiner, M.A. Roubidoux, and M.A. Helvie, "Computer-aided detection of masses in digital tomosynthesis mammography: combination of 3D and 2D detection information," Medical Imaging 2007: Computer-Aided Diagnosis, San Diego, CA, USA, **6514**, 651416-651416 (2007).

[22] G. Peters, S. Muller, S. Bernard, R. Iordache, and I. Bloch, "Reconstruction-independent 3D CAD for mass detection in digital breast tomosynthesis using fuzzy particles," Medical Imaging 2006: Image Processing, San Diego, CA, USA, **6144**, 61441Z-61410 (2006).

[23] H.P. Chan, J. Wei, B. Sahiner, E.A. Rafferty, T. Wu, M.A. Roubidoux, R.H. Moore, D.B. Kopans, L.M. Hadjiiski, and M.A. Helvie, "Computer-aided detection system for breast masses on digital tomosynthesis mammograms: preliminary experience." Radiology **237**, 1075-1080 (2005).

[24] S. Singh, G.D. Tourassi, and J.Y. Lo, "Breast mass detection in tomosynthesis projection images using information-theoretic similarity measures," SPIE Medical Imaging 2007: Computer-Aided Diagnosis, San Diego, CA, **6515** (2007).

[25] G.D. Tourassi and C.E. Floyd, Jr., "Knowledge-based detection of mammographic masses: analysis of the impact of database comprehensiveness," Medical Imaging 2005: PACS and Imaging Informatics, San Diego, CA, USA, **5748**, 399-406 (2005).

[26] G.D. Tourassi, B. Harrawood, S. Singh, and J.Y. Lo, "Information-theoretic CAD system in mammography: Entropy-based indexing for computational efficiency and robust performance." Med. Phys. **34**, 3193-3204 (2007).

[27] E. Samei, S.A. Stebbins, J.T. Dobbins, III, and J.Y. Lo, "Multiprojection Correlation Imaging for Improved Detection of Pulmonary Nodules." Am. J. Roentgenol. **188**, 1239-1245 (2007).

[28] B. Zheng, Y.H. Chang, and D. Gur, "Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis." Acad. Radiol. **2**, 959-966 (1995).

[29] B. Zheng, Y.H. Chang, and D. Gur, "Adaptive computer-aided diagnosis scheme of digitized mammograms." Acad. Radiol. **3**, 806-814 (1996).

[30] W.E. Polakowski, D.A. Cournoyer, S.K. Rogers, M.P. DeSimio, D.W. Ruck, J.W. Hoffmeister, and R.A. Raines, "Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency." IEEE Trans. Med. Imaging **16**, 811-819 (1997).

[31] Y. Chen, J.T. Dobbins, and J.Y. Lo, "Importance of point-by-point back projection correction for isocentric motion in digital breast tomosynthesis: Relevance to morphology of structures such as microcalcifications." Med. Phys. **34**, 3885-3892 (2007).

[32] M. Bissonnette, M. Hansroul, E. Masson, S. Savard, S. Cadieux, P. Warmoes, D. Gravel, J. Agopyan, B. Polischuk, W. Haerer, T. Mertelmeier, J.Y. Lo, Y. Chen, J.T. Dobbins Iii, J.L. Jesneck, and S. Singh, "Digital breast tomosynthesis using an amorphous selenium flat panel detector," Medical Imaging 2005: Physics of Medical Imaging, San Diego, CA, USA, **5745**, 529-540 (2005).

[33] T. Mertelmeier, J. Orman, W. Haerer, and M.K. Dudam, "Optimizing filtered backprojection reconstruction for a breast tomosynthesis prototype device," Medical Imaging 2006: Physics of Medical Imaging, San Diego, CA, USA, **6142**, 61420F-61412 (2006).

[34] T. Cover and J. Thomas, *Elements of information theory*. (Wiley-Interscience, 1991).

[35] K. Suzuki, S.G. Armato, 3rd, F. Li, S. Sone, and K. Doi, "Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography." Med. Phys. **30**, 1602-1617 (2003).

[36] K. Suzuki, H. Yoshida, J. Nappi, and A.H. Dachman, "Massive-training artificial neural network (MTANN) for reduction of false positives in computer-aided detection of polyps: Suppression of rectal tubes." Med. Phys. **33**, 3814-3824 (2006).

[37] K. Suzuki, J. Shiraishi, H. Abe, H. MacMahon, and K. Doi, "False-positive reduction in computer-aided diagnostic scheme for detecting nodules in chest radiographs by means of massive training artificial neural network1." Acad. Radiol. **12**, 191-201 (2005).

[38] K. Suzuki, K. Suzuki, L. Feng, S. Sone, and K.A.D.K. Doi, "Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network." Medical Imaging, IEEE Transactions on **24**, 1138-1150 (2005).

[39] G.D. Tourassi, B. Harrawood, S. Singh, J.Y. Lo, and C.E. Floyd, Jr., "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms." Med. Phys. **34**, 140-150 (2007).

[40] Y.H. Chang, L.A. Hardesty, C.M. Hakim, T.S. Chang, B. Zheng, W.F. Good, and D. Gur, "Knowledge-based computer-aided detection of masses on digitized mammograms: A preliminary assessment." Med. Phys. **28**, 455-461 (2001).

[41] G.D. Tourassi, R. Vargas-Voracek, J.D.M. Catarious, and J.C.E. Floyd, "Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information." Med. Phys. **30**, 2123-2130 (2003).

[42] A.O. Bilska-Wolak and C.E. FloydJr, "Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer." Phys. Med. Biol. **49**, 4219-4237 (2004).

[43] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*. (Chapman & Hall, New York, NY, 1993).

[44] D. Yates, D. Moore, and D. Starnes, *The Practice of Statistics*. (W H Freeman & Co, 2006).

[45] D.P. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data." Med. Phys. **16**, 561-568 (1989).

Effect of Similarity Metrics and ROI Sizes in Featureless Computer Aided Detection of Breast Masses in Tomosynthesis

Swatee Singh[1,2], Georgia D. Tourassi, Joseph Y. Lo[1,2,3]

[1]Duke Advanced Imaging Laboratories,
Duke University Medical Center
Durham, NC 27705
Tel. 919-684-1440, Fax 919-684-1491

[2]Department of Biomedical Engineering
Duke University
Durham, NC 27708-0281

[3]Department of Radiology, Medical Physics Graduate Program
Duke University Medical Center
Swatee.Singh@Duke.edu

Abstract. Tomosynthesis as a technique is being developed and studied with the goal of overcoming mammography's limitations due to overlying tissue. Various algorithms exist for tomosynthesis datasets including a novel Computer Aided Detection (CADe) algorithm using a featureless False Positive (FP) reduction stage. The goal of this project is to study the previously unexplored effects of variation of Region of Interest (ROI) sizes as well as the crucial similarity metrics for such a CADe algorithm's performance. Four datasets consisting of 1479 tomosynthesis ROIs were generated by a CADe algorithm from reconstructed volumes of one hundred subjects consisting of 4 different sizes – 128 x 128, 256 x 256, 512 x 512, and 1024 x 1024 pixels. Five different similarity metrics – (1) mutual information, (2) average conditional entropy, (3) joint entropy, (3) Jensen divergence and (4) average Kullback-Leibler divergence were used for the task of FP reduction using a leave-one-case-out sampling scheme. Mutual information and average conditional entropy were the best performing metrics with an Area Under Curve (AUC) of 0.88. Cross-bin measures performed consistently higher than those that rely on only marginal distributions. Also, for all metrics, the datatset consisting of 256 x 256 pixel ROIs gave the best performance. In conclusion, for the tomosynthesis dataset, cross-bin measures such as MI and average conditional entropy should be used over other metrics using a ROI size of 256 x 256 pixels.

Keywords: Tomosynthesis, 3D CAD, Computer Aided Detection, mammography, x-ray.

1 Background

Mammography remains the primary screening tool for breast cancer today. However, mammography has limitations because it takes two-dimensional images of a three-dimensional breast. Hence, overlying tissue can easily make it impossible to see underlying masses. Tomosynthesis, a limited angle cone-beam CT technique, has been proposed to overcome this shortcoming by providing radiologists reconstructed volumes of breasts to look at. These reconstructed volumes often consist of tens of slices, all of which the radiologist needs to look at thus potentially increasing their reading time per case. Given that nearly eighteen million women get their mammograms annually is US, this increase can have dramatic consequences on clinical workflow.

Many CADe algorithms exist for tomosynthesis datasets [1-7]. A novel approach for such CADe algorithms is to reduce False Positives (FPs) via featureless CADe using information theory principles [8-15]. Some research has been done using previously determined ROI sizes and similarity metrics for tomosynthesis datasets. However, there remains the potential of improving on these performances by optimization of these parameters.

2 Methods

Reconstructed breast volumes of one hundred subjects were used in this study consisting of 25 mass and 75 normal volumes. The average size of the lesions in this dataset was approximately 100 x 100 pixels (8.5 x 8.5 mm). Four distinct Region of Interest (ROI) datasets were extracted from these volumes consisting of 128 x 128, 256 x 256, 512 x 512, and 1024 x 1024 pixels (where pixel pitch was 85 microns). Each of these datasets consisted of 1479 computer generated tomosynthesis ROIs from reconstructed volumes that were extracted using an existing CADe algorithm.

For each of these datasets, False Positive (FP) reduction was done using principles of Information Theory wherein similarity metrics are computed between each ROI of unknown pathology to those in a database of pre-existing ROIs of known pathology. Such a database of ROIs consisting of known pathologies is often referred to as the Knowledge Base (KB). Many similarity metrics have been studied for the task of false positive reduction in mammographic ROIs. In this study we have investigated the following five metrics:

1. Mutual Information:

$$MI(X,Y) = \sum_{x}\sum_{y} P_{XY}(X,Y)\log_2\left(\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right) \quad \textbf{(1)}$$

2. Joint entropy:

$$Jo\mathrm{int}\,H = -\sum_{x}\sum_{y} p_{XY}(x,y)\log\left(p_{XY}(x,y)\right) \qquad (\mathbf{2})$$

3. Average conditional entropy:

$$\overline{conditional\_H} = \frac{H(x \mid y) + H(y \mid x)}{2} \qquad (\mathbf{3})$$

4. Jensen divergence:

$$JD(p,q) = \sum_{x}\left(q(x)\log\frac{2q(x)}{p(x)+q(x)} + p(x)\log\frac{2p(x)}{p(x)+q(x)} \right) \qquad (\mathbf{4})$$

5. Average Kullback-Leibler divergence:

$$avg\_divergence = \frac{D(q \parallel p) + D(p \parallel q)}{2} \qquad (\mathbf{5})$$

Where,

$$D(q \parallel p) = \sum_{x} q(x)\log\left(\frac{q(x)}{p(x)}\right) \qquad (\mathbf{6})$$

Where *X* and *Y* are the two images, $p(x,y)$ is their joint probability mass function, $p(x)$ and $p(y)$ are the marginal probability mass functions of *X* and *Y*, and *H(x|y)* and *H(y|x)* are the conditional entropies of the two images. All such metrics were then combined using a decision index to obtain a CADe score. These scores were then thresholded to yield ROC curves and the consequent AUCs and partial area under curve (pAUCs).

In some cases, it is the Contact Volume Editor that checks all the pdfs. In such cases, the authors are not involved in the checking phase.

3 Results

Results for all four datasets when each of the 5 metrics were individually evaluated are graphically represented in figure 1. For the best performing metrics, mutual information and average conditional entropy, the metrics along with their partial AUCs of > 0.90 sensitivity are listed in table 1. Figure 2 shows a human subject example with 2 reconstructed slices when using a classifier that works with 256x256 pixel ROIs and Conditional Entropy as a similarity metric. The true positive location is preserved during the false positive reduction stage, however 2 other false positives survive along with it as well.

ROI size	Mutual Information		Average conditional Entropy	
	AUC	pAUC	AUC	pAUC
128 x 128	0.61 +/- 0.02	0.08 +/- 0.03	0.62 +/- 0.02	0.08 +/- 0.03
256 x 256	0.88 +/- 0.02	0.49 +/- 0.09	0.88 +/- 0.02	0.48 +/- 0.09
512 x 512	0.70 +/- 0.04	0.27 +/- 0.04	0.66 +/- 0.04	0.10 +/- 0.03
1024 x 1024	0.58 +/- 0.03	0.01 +/- 0.04	0.57 +/- 0.03	0.01 +/- 0.04

Table 1. Various AUCs and pAUCs (AUC > 0.9) for the best performing metrics – mutual information and average conditional entropy

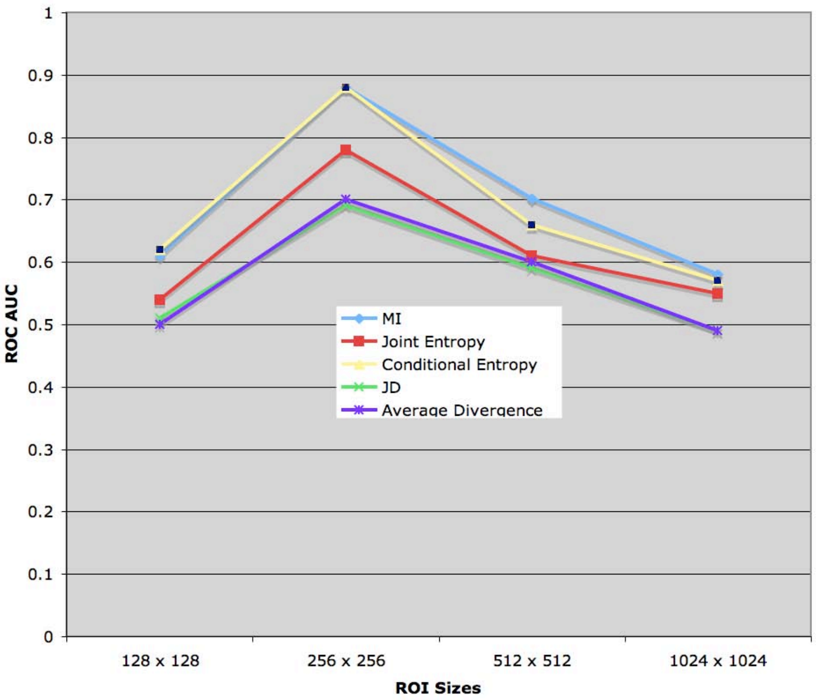


Fig. 1. Variation in ROC AUC for the four datasets being investigated consisting of varying ROI sizes for all 5 metrics being explored in this study.

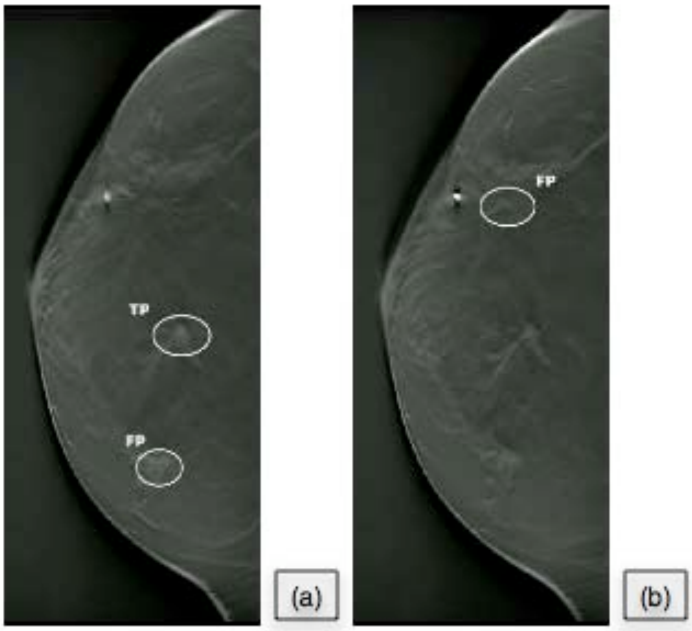


Fig. 2. Human subject example (a) Reconstructed slice with 1 true positive and 1 false positive each. (b) Reconstructed slice with another false positive within the same reconstructed volume. All results obtained when using a classifier that utilizes 256x256 pixel ROIs and Conditional Entropy as a similarity metric.

4 Discussion

Encouraging performance was obtained using the novel featureless CADe FP reduction scheme on tomosynthesis ROIs. The best performance across all measures was obtained using 256 x 256 pixel ROIs. Since the average lesion size in the dataset was 100 x 100 pixels, this reinforces the need to fully encompass the mass for optimal performance. Also, cross-bin measures that utilize joint probability distributions consistently outperformed measures that rely on only the marginal distributions for knowledge-based discrimination of masses from normal regions. Among cross-bin measures that incorporate information from non-corresponding bins, mutual information and conditional entropy outperformed joint entropy. It is possible that cross-bin measures are more capable at capturing some aspects of visual content more effectively than bin-by-bin measures. When the ROI size reaches 1024 x 1024 pixels, the performance of all similarity metrics approaches chance.

However, a drawback of information theory based techniques is that the localized spatial relationships among the image pixels are lost when working with image histograms. This limitation has been addressed before in the context of image registration. It has been proposed that taking into account the neighborhood of regions of corresponding image pixels may be a more effective strategy. In on-going work, these results will be applied to full image volumes to assess free-response ROC results.

References

1. Reiser, I., Nishikawa, R.M., Giger, M.L., Wu, T., Rafferty, E.A., Moore, R., Kopans, D.B.: Computerized mass detection for digital breast tomosynthesis directly from the projection images. Med. Phys. **33** (2006) 482-491
2. Reiser, I.S., Sidky, E.Y., Giger, M.L., Nishikawa, R.M., Rafferty, E.A., Kopans, D.B., Moore, R., Wu, T.: A reconstruction-independent method for computerized mass detection in digital tomosynthesis images of the breast. Medical Imaging 2004: Image Processing, Vol. 5370. SPIE, San Diego, CA, USA (2004) 833-838
3. Chan, H.-P., Wei, J., Zhang, Y., Moore, R.H., Kopans, D.B., Hadjiiski, L., Sahiner, B., Roubidoux, M.A., Helvie, M.A.: Computer-aided detection of masses in digital tomosynthesis mammography: combination of 3D and 2D detection information. Medical Imaging 2007: Computer-Aided Diagnosis, Vol. 6514. SPIE, San Diego, CA, USA (2007) 651416-651416
4. Wheeler, F.W., Perera, A.G.A., Claus, B.E., Muller, S.L., Peters, G., Kaufhold, J.P.: Micro-calcification detection in digital tomosynthesis mammography. Medical Imaging 2006: Image Processing, Vol. 6144. SPIE, San Diego, CA, USA (2006) 614420-614412
5. Bernard, S., Muller, S., Peters, G., Iordache, R.: Fast microcalcification detection on digital breast tomosynthesis datasets. Medical Imaging 2007: Computer-Aided Diagnosis, Vol. 6514. SPIE, San Diego, CA, USA (2007) 65141X-65148
6. Peters, G., Muller, S., Grosjean, B., Bernard, S., Bloch, I.: A hybrid active contour model for mass detection in digital breast tomosynthesis. Medical Imaging 2007: Computer-Aided Diagnosis, Vol. 6514. SPIE, San Diego, CA, USA (2007) 65141V-65111
7. Peters, G., Muller, S., Bernard, S., Iordache, R., Bloch, I.: Reconstruction-independent 3D CAD for mass detection in digital breast tomosynthesis using fuzzy particles. Medical Imaging 2006: Image Processing, Vol. 6144. SPIE, San Diego, CA, USA (2006) 61441Z-61410
8. Tourassi, G.D., Vargas-Voracek, R., Catarious, D.M., Jr: Computer-assisited detection of mammographic masses: a template matching scheme based on mutual information. Med. Phys. **30** (2003) 2123-2130
9. Tourassi, G.D., Harrawood, B., Singh, S., Lo, J.Y.: Information-theoretic CAD system in mammography: Entropy-based indexing for computational efficiency and robust performance. Med. Phys. **34** (2007) 3193-3204
10. Tourassi, G.D., Harrawood, B., Singh, S., Lo, J.Y., Floyd, C.E., Jr.: Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms. Med. Phys. **34** (2007) 140-150
11. Tourassi, G.D., Vargas-Voracek, R., Catarious, D.M., Jr., Floyd, C.E., Jr.: Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information. Med. Phys. **30** (2003) 2123-2130

12. Tourassi, G.D., Floyd, C.E., Jr.: Knowledge-based detection of mammographic masses: analysis of the impact of database comprehensiveness. Medical Imaging 2005: PACS and Imaging Informatics, Vol. 5748. SPIE, San Diego, CA, USA (2005) 399-406

13. Singh, S., Tourassi, G.D., Lo, J.Y.: Breast mass detection in tomosynthesis projection images using information-theoretic similarity measures. In: Giger, M.L., Karssemeijer, N. (eds.): SPIE Medical Imaging 2007: Computer-Aided Diagnosis, Vol. 6515, San Diego, CA (2007)

14. Ike Iii, R.C., Singh, S., Harrawood, B., Tourassi, G.D.: Effect of ROI size on the performance of an information-theoretic CAD system in mammography: multi-size fusion analysis. Medical Imaging 2008: Computer-Aided Diagnosis, Vol. 6915. SPIE, San Diego, CA, USA (2008) 691527-691527

15. Singh, S., Tourassi, G.D., Chawla, A.S., Saunders, R.S., Samei, E., Lo, J.Y.: Computer-aided detection of breast masses in tomosynthesis reconstructed volumes using information-theoretic similarity measures. Medical Imaging 2008: Computer-Aided Diagnosis, Vol. 6915. SPIE, San Diego, CA, USA (2008) 691505-691508

Evaluating the Effect of Image Preprocessing on an Information-Theoretic CAD System in Mammography[1]

Georgia D. Tourassi, PhD, Robert Ike III, BS, Swatee Singh, BS, Brian Harrawood, BS

Rationale and Objectives. In our earlier studies, we reported an evidence-based computer-assisted decision (CAD) system for location-specific interrogation of mammograms. A content-based image retrieval framework with information theoretic (IT) similarity measures serves as the foundation for this system. Specifically, the normalized mutual information (NMI) was shown to be the most effective similarity measure for reduction of false-positive marks generated by other prescreening mass detection schemes. The objective of this work was to investigate the importance of image filtering as a possible preprocessing step in our IT-CAD system.

Materials and Methods. Different filters were applied, each one aiming to compensate for known limitations of the NMI similarity measure. The study was based on a region-of-interest database that included true masses and false-positive regions from digitized mammograms.

Results. Receiver-operating characteristics (ROC) analysis showed that IT-CAD is affected slightly by image filtering. Modest, yet statistically significant, performance gain was observed with median filtering (overall ROC area index A_z improved from 0.78 to 0.82). However, Gabor filtering improved performance for the high-sensitivity portion of the ROC curve where a typical false-positive reduction scheme should operate (partial ROC area index ${}_{0.90}A_z$ improved from 0.33 to 0.37). Fusion of IT-CAD decisions from different filtering schemes markedly improved performance ($A_z = 0.90$ and ${}_{0.90}A_z = 0.55$). At 95% sensitivity, the system's specificity improved by 36.6%.

Conclusions. Additional improvement in false-positive reduction can be achieved by incorporating image filtering as a preprocessing step in our IT-CAD system.

Key Words. CAD; mammography; image processing; information theory.

© AUR, 2008

Despite advances in treatment, breast cancer remains the second leading cause of cancer death in women (1). The role of screening mammography in the battle against breast cancer is well established; women with malignancies detected at an early stage have a significantly better prognosis (2). However, it is also recognized that the diagnostic interpretation of mammograms continues to be challenging for radiologists with a documented 20% false-negative rate (3–6).

The clinical significance of early breast cancer diagnosis and the higher than desired false-negative rate of screening mammography have motivated the development of computer-aided detection (CADe) systems for decision support. These systems typically involve a hierarchical approach, first applying elaborate image preprocessing steps to enhance suspicious structures in the image and then employing morphologic and textural analysis to bet-

Acad Radiol 2008; 15:626–634

[1] Digital Advanced Imaging Laboratories, Department of Radiology (G.D.T., R.I., S.S., B.H.), Medical Physics Program (R.I.), and Department of Biomedical Engineering (S.S.), Duke University Medical Center, 2424 Erwin Road, Hock Plaza, Suite 302, Durham, NC 27710. Received November 4, 2007; accepted December 12, 2007. Supported by grant R01 CA101911 from the National Cancer Institute, Bethesda, MD. **Address correspondence to:** G.D.T. e-mail: georgia.tourassi@duke.edu

© AUR, 2008

doi:10.1016/j.acra.2007.12.013

ter classify these structures between true abnormalities and false positives. Detailed reviews of image processing techniques for mammographic image analysis and related CADe systems can be found elsewhere (7–10). In addition, several CADe systems are already available commercially for both screen film mammography and full-field digital mammography (7). Although their true clinical impact is often debated (11–19), the scientific community continues to work toward improving the diagnostic performance and clinical integration of CADe technology. Ongoing CADe research efforts focus mainly on the reduction of false-positive computer marks as well as improving the detection rate of breast masses.

In our earlier studies, we presented a knowledge-based CADe system for breast mass detection in screening mammograms (20–22). The system is interactive and is designed to operate as a second opinion for mammographic locations that are deemed suspicious of containing breast masses. These suspicious locations are areas that attract the radiologist's attention or are marked as suspicious by other automated mass detection schemes. Thus, our system is designed for location-specific interrogation of mammograms. The interrogation relies on a database of mass and normal examples with known ground truth. These examples serve as the knowledge database. Basically, the system compares the query location with the knowledge examples. The comparison is performed using featureless, information-theoretic (IT) similarity measures (21). Such measures are based on the concept of image entropy (23), and they are calculated directly from the image pixel intensity values. Although we explored various IT measures, our IT-CADe system using either mutual information (MI) or its normalized version (NMI) was shown to be the most effective (21,22).

The original IT-CADe prototype relied on raw image data without any preprocessing. Our present study reports on the effect of image preprocessing on the overall diagnostic performance of this system. Medical image registration studies using mutual information suggest that minimal preprocessing often improves image registration performance (24). Consequently, in this study, we explored the effect of various preprocessing image filters on our own IT-CADe system. The selection of each preprocessing filter targeted known limitations of the mutual information similarity measure. The resultant performance of the modified IT-CADe system was compared with that previously reported without the preprocessing filtering step. Such direct comparison is necessary to test the hypothesis that image preprocessing contributes to further improvement of the IT-CADe performance.

MATERIALS AND METHODS

Materials

Database.—The image database used in this study has been previously described in detail (20,21). Because the present study builds on a previously presented system, it is essential to demonstrate any incremental improvement using the same database. Here is a summary description of this database.

All mammographic cases were selected from the Digital Database of Screening Mammography (DDSM) (25). These mammograms were scanned with a Lumisys scanner (Sunnyvale, CA) at 50 μm/pixel and a bit depth of 12. There were 583 mammograms in total; 296 containing biopsy-proven malignant masses, 185 containing benign masses proven either by biopsy or additional imaging, and 82 normal mammograms. The available database was divided into two sets. One set contained 483 DDSM cases (256 cancer, 145 benign, 62 normal) and served as the knowledge database. The second set contained the remaining 100 cases (40 malignant, 40 benign, 20 normal) and served as the test database. Note that the test database was reserved from the beginning of our IT-CADe research efforts (before this study) to serve for final validation. The selection criteria were such that the test database represents a balanced mix of cases from all available DDSM/Lumisys volumes. The database did not contain any "benign-without-callback" cases, because these are considered easy cases to diagnose.

From each case, a 512 × 512 pixel region of interest (ROI) was extracted around the known location of any true mass present in the case. The mass locations are provided in the DDSM truth files. Dataset 1 (ie, the knowledge database) contained 1,820 ROIs. Of those, 489 depicted a malignant mass, 412 depicted a benign mass, and 919 were normal. The mass ROIs serve as the system's knowledge foundation of typical mass examples. The normal ROIs were initially selected from normal mammographic cases by randomly sampling the breast region. Such normal ROIs are essential to establish a knowledge foundation of normal breast parenchyma. Because the normal cases were few compared to mass cases in the DDSM/Lumisys set, normal ROIs were also extracted from abnormal cases, but only from imaged breasts that did not contain any physician annotations in either mammographic view.

In addition, 512 × 512 pixel ROIs were extracted around the known mass locations in the test database. There were 44 malignant mass ROIs and 40 benign mass ROIs in dataset 2. In addition, 399 ROIs were extracted around mammographic locations marked as suspicious by a feature-based CADe system developed before in our laboratory (26,27). Therefore, dataset 2 contained 483 ROIs in total. These ROIs served as queries to our IT-CADe system to determine whether the system can provide effective false-positive reduction.

Overview of the IT-CADe system.—The prototype IT-CADe system offers an evidence-based second opinion regarding the presence of a possible mass in any mammographic location that is indicated by the CADe user. The basic IT-CADe system combines principles from content-based image retrieval and case-based reasoning. When an unknown query case is presented for evaluation to the system, the system compares the query to all known cases stored in the knowledge database. Similar cases are retrieved and are used to make a prediction regarding the query case. The retrieval process relies on IT measures. Such measures include mutual information, joint entropy, and Kullback-Leibler divergence (23). Generally, these similarity measures are calculated using the image pixel intensity values directly, not any image features. The underlying assumption is that the co-occurrence of the intensity values in the two images is maximized when the images match well. The IT measures use the concept of entropy to measure the co-occurrence of pixel values (23).

Our previous publications (20,21,22,28) addressed issues related to the composition of the knowledge database, the case retrieval process, the construction of the decision index, and the effect of the similarity metric. Based on our previous findings, the prototype system operates as follows. First, a 512 × 512 pixel ROI is extracted around the suspicious mammographic location indicated by the CADe user or marked by another detection algorithm. The ROI serves as the query case for the system. Then the ROI is compared to all examples (or templates) stored in the system's knowledge database. Examples with similar entropy as that of the query are quickly identified using a previously presented entropy-based indexing scheme (22). The entropy-based indexing scheme serves as a search mechanism to sort through the knowledge database fast and identify the stored examples that are more relevant to the specific query. Then detailed pairwise comparisons are performed between the query (Q) and each relevant knowledge example (or template [T]). This detailed comparison is based on the NMI similarity measure (22). NMI captures the statistical dependence between two images and it is always bounded between 0 and 1. A value of 1 suggests perfect match between the query case Q and the stored template T. In contrast, a value of 0 indicates no statistical dependence or shared information between the two cases. Some studies in image registration have shown that NMI is often more successful and robust than MI (24,29,30). Although our previous studies showed that both MI and NMI are equally effective in IT-CADe (21), the bounded nature of NMI makes it a more attractive option. Finally, a decision index is calculated measuring how well the query case matches on average the retrieved mass templates compared to the retrieved normal templates. In a clinical setting, an optimal threshold needs to be determined for the final decision. If the query's decision index exceeds the threshold value, then the query mammographic location is marked as a true mass. Otherwise, the query location is marked as normal.

METHODS

Preprocessing filters.—To test the effect of image preprocessing on the system performance, we applied several different filters. These filters were selected to compensate for potential limitations of the NMI similarity measure, such as lower robustness in the presence of noise, lack of spatial information, and questionable perceptual relevance. Specifically, five different filters were investigated. Two of them were popular denoising filters, namely the median and the adaptive Wiener filter. The third choice was a perceptually driven Gabor filter. Finally, two texture filters were also investigated, an entropy-based and a localized standard deviation filter. Image filtering was performed using the MATLAB programming environment (The MathWorks, Inc, Natick, MA).

(a) Median filter: Several image registrations studies suggested that noise reduction techniques are essential for more accurate MI-based image registration (24,31). We have explored the same issue for our IT-CADe system. Specifically, we applied median filtering before the calculation of the similarity measures. Median filtering is a standard noise reduction technique. Furthermore, it is a reasonable preprocessing step for mass detection because it preserves the edge information of suspicious areas while reducing noise (32). Median filters with different size kernels were explored (3 × 3, 5 × 5, 7 × 7, 9 × 9, 11 × 11, 15 × 15, 21 × 21 pixels) for the task.

(b) Adaptive Wiener filter: Similar to the median filter, the adaptive Wiener filter (33) was applied for denoising tailored on statistics estimated from the local neighborhood of each image pixel. The amount of smoothing performed by the filter depends on the local image mean and variance around the pixel of interest. The Wiener filter is a popular linear filter, but its adaptive implementation better preserves the high-frequency parts of the image. The same size kernel sizes were explored as with the median filter.

(c) Gabor filter: Another promising filter for image denoising and texture analysis is the Gabor filter (34). This type of multichannel filtering is considered an excellent preprocessing choice for image registration (35–37) because of its perceptual relevance (38). Specifically, it has been shown that Gabor filters model the spatial frequency and orientation responses of simple cells in the primary visual cortex (39,40). The Gabor representation has been shown to be optimal in the sense of minimizing the joint two-dimensional uncertainty in space and frequency (41). Because the Gabor filter bank is derived from a wavelet basis with dilations and orientations, they are essentially band-pass filters.

A two-dimensional symmetric Gabor filter was implemented as described in Eq 1:

$$f(x,y) = e^{\left\{-\frac{1}{2}\left[\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right]\right\}} \cdot \cos(2\pi\mu_0(x\cos\theta + y\sin\theta)) \quad (1)$$

where μ_0 is the frequency of a sinusoidal plane, θ is the orientation, and σ_x and σ_y are standard deviations (or spatial spread) of the two-dimensional Gaussian envelope (42). An octave bandwidth of 1 was used in our study because past psychophysical studies have confirmed that an octave bandwidth of 1 is a reasonably good estimate of the human eye when tuned to a frequency (43). Central frequencies of 0.5, 1, 2, 4, 8, 16, and 32 cycles/degree with orientations at 0°, 45°, 90°, and 135° were used in this study.

(d) Entropy-based filtering: Because NMI is calculated using only intensity information of corresponding pixels between two images, it has an inherent limitation. It ignores possible relationships between neighboring pixels. Because image texture is typically captured by such relationships, NMI ignores a potentially critical diagnostic component. To address this limitation, we investigated an entropy-based filter as a preprocessing step for all images. The filter replaces the intensity value of each image pixel with a new value that captures the local image entropy around the pixel (44). Thus, each pixel value is replaced with a new value that contains localized textural information. This filter was implemented using the *entropyfilt* function in the MATLAB Image Processing Toolbox. Multiscale analysis was investigated repeating this filtering step at several scales by varying the neighborhood size of the entropy-based filter (3 × 3, 5 × 5, 7 × 7, 9 × 9, 11 × 11, 15 × 15, and 21 × 21 pixels).

(e) Standard deviation filter: Similar to the entropy-based filter, the standard deviation filter replaces each pixel value of the grayscale image with the local standard deviation of a neighborhood around the pixel of interest. This preprocessing filter was implemented using the *stdfilt* function of the MATLAB Image Processing Toolbox; it was also evaluated for variable size neighborhoods as the entropy-based filter.

Figure 1 shows a representative, unprocessed ROI depicting a malignant mass along with its filtered versions using the following filters: median, locally adaptive Wiener, Gabor, entropy based, and standard deviation based.

Evaluation Methods

Both datasets 1 and 2 were preprocessed using the previously described filters. For each separate filter, the IT-CADe system was tested using dataset 1 as the knowledge database and dataset 2 as the test bed for the discrimination of true masses from false-positive findings. Detection performance was evaluated with receiver-operating characteristic (ROC) analysis (45). ROC curves were fitted with the ROCKIT software, available from Charles Metz at the University of Chicago. The overall ROC area A_z and the partial ROC area ${}_{0.90}A_z$ were used as the reported performance indices. Although A_z is the most common performance index for binary diagnostic tasks taking into account all possible decision thresholds, the partial ROC area index summarizes the detection performance for decision thresholds corresponding only to

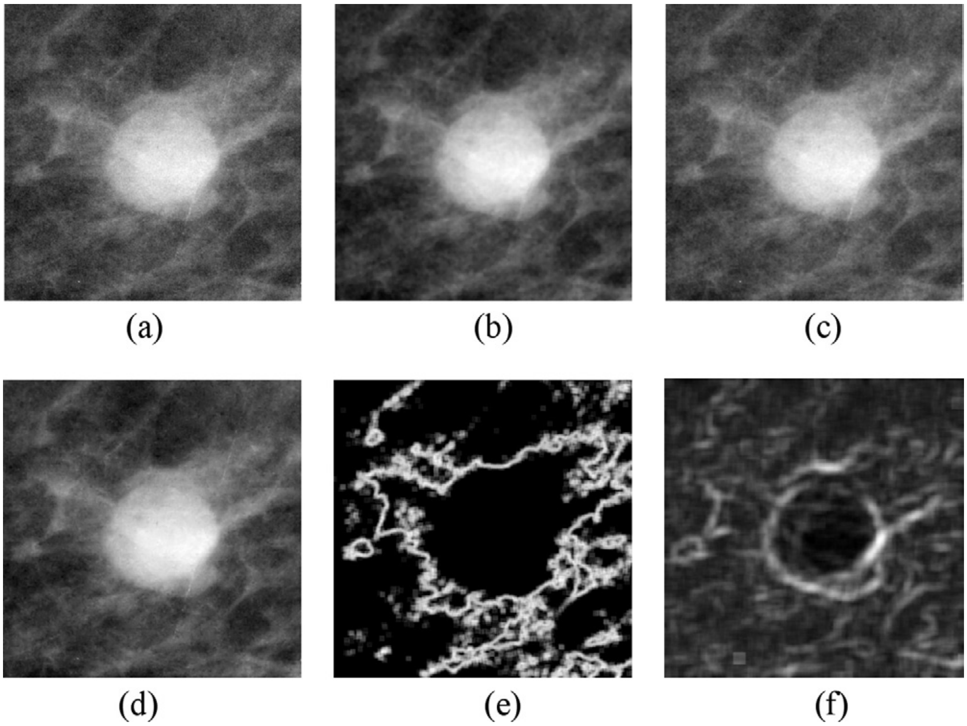


Figure 1. Example region of interest (ROI) depicting a malignant mass **(a)** unprocessed, and processed with the following filters: **(b)** 3 × 3 median, **(c)** 3 × 3 adaptive Wiener, **(d)** Gabor, **(e)** 9 × 9 entropy-based, and **(f)** 21 × 21 standard deviation-based.

the high-sensitivity portion (>90%) (46). For our study, $_{0.90}A_z$ is certainly a more appropriate performance index because any false-positive reduction scheme is expected to perform at a high cancer detection rate for a clinically effective cancer screening CADe system.

RESULTS

Effect of Image Filter

First, the effect of the kernel size on IT-CADe performance was investigated carefully for the median, Wiener, entropy-based, and standard deviation-based filters. Figure 2 shows the corresponding ROC area index A_z (Fig 2a) and partial ROC area index $_{0.90}A_z$ (Fig 2b) for all kernel sizes considered. For the median filter, the 3 × 3 kernel resulted in the highest ROC performance with the 5 × 5 kernel producing a slightly (yet not statistically significantly) lower performance. As the kernel size of the median filter increased, the performance of the system steadily decreased. This result was expected because of the resulting oversmoothing of the images. The kernel size of the Wiener filter had minimal impact on the system's performance, at least for the size range evaluated in this study. Similar to the median filter, the neighborhood size of the texture filters also affected system performance. Performance peaked for the 9 × 9 neighborhood size with the entropy-based filter. The improvement was statistically significant compared to all other neighborhood sizes with the exception of the 11 × 11 neighborhood, where the difference was borderline significant (two-tailed *P* value of .05 for the partial ROC area index). For the standard deviation filter, ROC performance peaked for the 21 × 21 neighborhood size, but it was not significantly better compared to the other neighborhood sizes. Further increase of the neighborhood size resulted in a severe decrease of the system's performance.

Table 1 summarizes the results of this study for all preprocessing scenarios considered and shows the performance indices for the IT-CADe system depending on the image preprocessing scheme. For simplicity, Table 1 shows only the system performance for each preprocessing filter operating with its best performing kernel size. As a point of reference, the table also includes the performance of the original IT-CADe system without any image preprocessing ("none") and reports the specificity achieved by the system at 95% detection rate for masses.

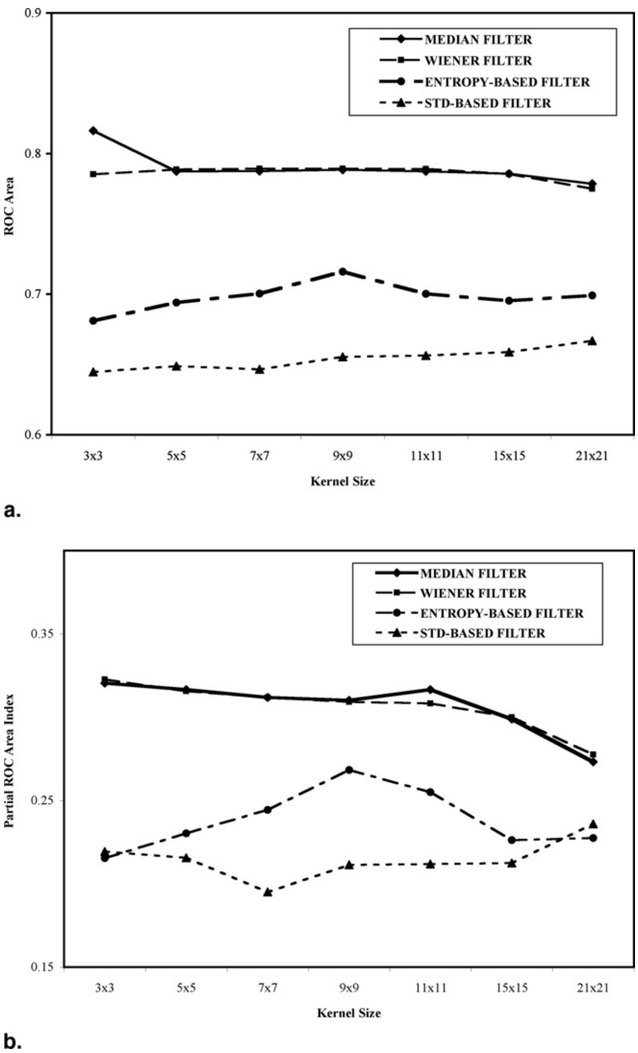


Figure 2. Effect of the filter kernel size on the receiver-operating characteristic (ROC) performance of the information theoretic–computer-aided detection (IT-CADe) system with respect to the **(a)** overall ROC area index and the **(b)** partial ROC area index for the high-sensitivity (>90%) portion of the ROC curve.

Table 1 highlights several interesting trends. Overall, both texture filters resulted in a dramatic decline of the IT-CADe diagnostic performance with respect to all performance indices. The decline of the ROC area index was significant for both the entropy-based and standard deviation–based filters (two-tailed *P* value $<$ .001). With respect to the partial area index, the decline was borderline significant for the standard deviation filter (two-tailed *P* value of .05), but not significant for the entropy filter (two-tailed *P* value of .12). These results suggest that the texture filters investigated in this study are not appropriate choices if they are to be used as an independent preprocessing step. However, because they capture textural information, such filters have potentially incremental diagnostic value.

With respect to the overall ROC area index, the median filter resulted in a statistically significant improvement of the diagnostic performance. The area index increased from 0.78 to 0.82 (two-tailed *P* value of .01). However, such improvement was not observed with respect to the partial ROC area index. Actually, the partial ROC area index declined slightly from 0.33 to 0.32 after median filtering (two-tailed *P* value of .20). The Wiener filter resulted in similar performance of the IT-CADe system without any preprocessing (two-tailed *P* values of .53 and .29 for the ROC and partial ROC area indices, respectively). Finally, Gabor filtering produced a notable improvement of the partial ROC area index (from 0.33 to 0.37); however, this improvement did not reach statistical significance (two-tailed *P* value of .12).

With respect to specificity at 95% mass detection rate, Gabor filtering was the most effective. The system achieved 34.1% specificity when including Gabor filtering as a preprocessing step compared to 31.3% specificity without filtering. This result represents a 9% improvement in system specificity.

IT-CADe Fusion

Although no filter emerged as a clearly superior choice, it is possible that a multifilter fusion approach may be more effective. To test this possibility, we constructed a linear classifier that combined the predictions of the IT-CADe systems (each operating with a different preprocessing step) into one comprehensive decision. The underlying hypothesis is that fusing the IT-CADe outputs based on multiple, complementary preprocessing filters may be superior to any one of the filters alone.

Specifically, linear classifiers were built combining the filter-specific IT-CADe outputs. For a given query ROI, the continuous decision indices of the filter-specific IT-CADe systems served as inputs to the fusion classifier. Thus, the fusion CAD system relied on stacked generalization where the level 0 classifiers were the knowledge-based, filter-specific IT-CADe systems and the level 1 combiner was a trainable linear classifier. We performed an exhaustive search, building a linear classifier for every possible combination of "filters" (ie, filtered IT-CADe outputs). With six different filters considered, there were 57 possible combinations (ie, 15 combinations merging the IT-CADe outputs of only two different filters at a time, 20 combinations merging three different filters, 15 combinations merging four different filters, six combina-

Table 1
Effect of Image Filtering as a Preprocessing Step on the Performance of the IT-CADe System for the Detection of Masses in Screening Mammograms

Preprocessing Filter	A_z (± σ)	$_{0.90}A_z$ (± σ)	Specificity at 95% Sensitivity
None	0.778 ± 0.025	0.326 ± 0.055	31.3% (125/399)
Median (3 × 3)	0.816 ± 0.025	0.320 ± 0.065	29.6% (118/399)
Wiener (3 × 3)	0.785 ± 0.026	0.323 ± 0.057	31.1% (124/399)
Gabor	0.783 ± 0.024	0.368 ± 0.053	34.1% (136/399)
Entropy (9 × 9)	0.706 ± 0.028	0.268 ± 0.046	27.6% (110/399)
Standard deviation (21 × 21)	0.667 ± 0.028	0.236 ± 0.042	23.8% (95/399)

IT-CADe: information theoretic–computer-aided detection.

Table 2
Performance of Linear Discriminant Analysis Decision Models that Combine the Filter-Specific IT-CADe Outputs

LDA	A_z (± σ)	$_{0.90}A_z$ (± σ)	Specificity at 95% Sensitivity
2 filters: (M, W)	0.884 ± 0.019	0.517 ± 0.067	49.4% (197/399)
3 filters: (M, W, STD)	0.893 ± 0.018	0.523 ± 0.070	47.4% (189/399)
4 filters: (M, W, H, STD)	0.896 ± 0.017	0.535 ± 0.067	48.3% (193/399)
5 filters: (M, W, G, STD, UN)	0.897 ± 0.018	0.549 ± 0.067	49.9% (199/399)
ALL: (M, W, G, H, STD, UN)	0.898 ± 0.018	0.548 ± 0.068	50.4% (201/399)

LDA: linear discriminant analysis; IT-CADe: information theoretic–computer-aided detection; UN: unprocessed; M: median; W: adaptive Wiener; G: Gabor; H: entropy-based; STD: standard deviation based.

Different LDA models were built for each possible combination of filtering options. The table shows which combinations emerged as the superior ones depending on the number of inputs allowed in the LDA model.

tions merging five filters, and one combination including all six filtered IT-CADe outputs). Thus, 57 different linear classifiers were built. These classifiers were evaluated using leave-one out sampling on dataset 2 because the clinical focus was on differentiating masses from false positives. Furthermore, in our previous experiments we observed that leave one out is an appropriate data-handling scheme when stacking knowledge-based IT-CADe systems with a simple combiner such as a linear classifier (47). These experiments were performed using the R software package (48,49).

Table 2 highlights some of the most interesting trends of the IT-CADe fusion experiment. Specifically, the table shows which combination produced the best performing fusion classifier when the number of input filters is restricted (eg, only two filters, only three filters). Overall, the fusion experiment showed that the synergistic approach of the linear classifiers using information from IT-CADe with different preprocessing schemes resulted in statistically significantly better performance compared to the original IT-CADe system with respect to all performance metrics. Fusing all six filtering schemes produced the best results (A_z = 0.90, $_{0.90}A_z$ = 0.55). At a fixed 95% mass sensitivity rate, 76 additional false-positive queries were correctly identified by the fusion system. This represents a 61% specificity improvement over the original IT-CADe system. However, combining the decisions of the IT-CADe system operating with only two preprocessing steps, namely the median and the adaptive Wiener filters, produced similar results (A_z = 0.88, $_{0.90}A_z$ = 0.52, and 57.8% specificity improvement at a fixed 95% mass detection rate). This performance was not significantly lower than the best reported one using all six filters. Compared to the best performing median filter, the two-filter fusion linear discriminant analysis increased specificity from 31.3% to 49.4%. Similarly, a 44.9% specificity improvement was observed over the system operating with the best performing Gabor filter (specificity increased from 34.1% to 49.4%). Although several combinations of three, four, and five filters resulted in incremental performance improvements (shown in Table 2), none of these was statistically significant compared to combining just the median and Wiener filters. This result suggests that the added complexity of additional filters is not justified.

DISCUSSION

The general concept of building and mining knowledge databases of imaging data in radiology is becoming increasingly relevant. In the digital era, it is important to capitalize on the growing number and variety of mammograms that are continuously acquired and stored. Our interactive, knowledge-based CADe system for location-specific interrogation of mammograms has such capacity without needing additional training of its decision-making module every time a new case is added to its knowledge database. Furthermore, our featureless approach for case similarity assessment eliminates any concerns regarding careful selection, extraction, and merging of image features for decision making. This is a particularly attractive property that facilitates easier knowledge transfer across databases (eg, mammograms acquired with different systems or digitized with different digitizers), as we have shown in previous studies (28,50).

In this study, we presented a range of image filtering techniques as potential preprocessing steps in an attempt to improve the performance of our IT-CADe system. The filters were selected so that they complemented the similarity metric in our IT-CADe system. Because normalized mutual information is sensitive to image noise, a smoothing median filter and an adaptive Wiener filter were considered as promising preprocessing steps. In addition, two texture-based filters were considered. Because NMI does not capture localized textural information, an entropy-based filter and a standard deviation-based filter were applied to quantify the local texture in the image. The final choice was a Gabor filter optimized according to the human perception system. This comparative study focused on the false-positive reduction task because such task still represents one of the major challenges of existing CADe systems in mammography.

Our study was restricted to a small but diverse group of preprocessing filters. Overall, no particular filter emerged as the superior choice. Although median filtering resulted in significantly better performance with respect to the overall ROC area, Gabor filtering demonstrated superior performance for the clinically critical, high-sensitivity portion of the ROC curve. However, the improvement did not reach statistical significance. The entropy-based and standard deviation–based texture filters were the only ones that deteriorated the diagnostic performance of the IT-CADe system. It should be noted that other texture-based filters that are better tailored to the clinical task could be potentially more successful. Finally, integrating all filters with a linear classifier achieved dramatic improvement with respect to all performance indices. These results highlight the significance of image preprocessing for our IT CADe system, especially when a fusion approach is considered in which the filters are complementary in nature.

In conclusion, image preprocessing through carefully tailored filters should be investigated as a promising strategy to improve substantially upon the performance of our CADe system. Moreover, advanced fusion strategies that incorporate CADe decisions based on complementary preprocessing steps hold the most promise for providing even further improvements.

REFERENCES

1. American Cancer Society. American Cancer Society: cancer facts and figures 2002. Atlanta, GA: American Cancer Society, 2002.
2. Tabar L, Vitak B, Chen HH, et al. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. Cancer 2002; 91:1724–1731.
3. Beam CA, Conant EF, Sickles EA. Factors affecting radiologist inconsistency in screening mammography. Acad Radiol 2002; 9:531–540.
4. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology 1992; 184:613–617.
5. Birdwell RL, Ikeda DM, O'Shaughnessy KF, et al. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology 2001; 219:192–202.
6. Yankaskas BC, Schell MJ, Bird RE, et al. Reassessment of breast cancers missed during routine screening mammography: a community-based study. AJR Am J Roentgenol 2001; 177:535–541.
7. Sampat MP, Markey MK, Bovik AC. Computer-aided detection and diagnosis in mammography. In: Bovik AC, ed. Handbook of image and video processing, 2nd ed. NY: Academic Press 2005; 1195–1217.
8. Thangavel K, Karnan M, Sivakumar R, et al. Automatic detection of microcalcification in mammograms—a review. ICGST-GVIP J 2005; 5:31–61.
9. Highnam RP, Brady JM. Mammographic image analysis. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1999.
10. Fitzpatrick JM, Sonka M. Handbook of medical imaging, vol. 2. Medical image processing and analysis. Bellingham, WA: SPIE Press, 2000.
11. Burhenne LJW, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000; 215:554–562.
12. Brem RF, Baum J, Lechner M, et al. Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. AJR Am J Roentgenol 2003; 81:687–693.
13. Gur D, Sumkin H, Rockette HE, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. J Natl Cancer Inst 2004; 96:185–190.
14. Taylor P, Given-Wilson RM. Evaluation of computer-aided detection (CAD) devices. Br J Radiol 2005; 78:26–30.
15. Birdwell RL, Bandodkar P, Ikeda DM. Computer-aided detection with screening mammography in a university hospital setting. Radiology 2005; 236:451–457.
16. Morton MJ, Whaley DH, Brandt KR, et al. Screening mammograms: Interpretation with computer-aided detection—prospective evaluation. Radiology 2006; 239:375–383.

17. Krupinski EA. Computer-aided detection in clinical environment: benefits and challenges for radiologists. Radiology 2004; 231:7–9.
18. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007; 356:1399–1409.
19. Hall FM. Breast imaging and computer-aided detection. N Engl J Med 2007; 356:1464–1466.
20. Tourassi GD, Vargas-Voracek R, Floyd CE, Jr. Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information. Med Phys 2003; 30:2123–2139.
21. Tourassi GD, Harrawood B, Singh S, et al. Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms. Med Phys 2007; 34:140–150.
22. Tourassi GD, Harrawood B, Singh S, et al. Information-theoretic CAD system in mammography: entropy-based indexing for computational efficiency and robust performance. Med Phys 2007; 34:3193–3204.
23. Cover TM, Thomas JA. Elements of information theory. New York: John Wiley & Sons, 1991.
24. Pluim JPW, Maintz JBA, Viergever MA. Mutual-information-based registration of medical images: a survey. IEEE Trans Med Imag 2003; 22: 986–1004.
25. Heath M, Bowyer K, Kopans D, et al. Current status of the digital database for screening mammography. In: Digital mammography. Kluwer Academic Publishers, 1998. Available online at: http://marathon.csee.usf.edu/Mammography/Database.html. Accessed January 7, 2008.
26. Catarious DM, Baydush AH, Floyd CE, Jr. A mammographic mass CAD system incorporating features from shape, fractal, and channelized Hotelling observer measurements: preliminary results. SPIE 2003; 5032: 111–119.
27. Catarious DM, Baydush AH, Floyd CE, Jr. Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system. Med Phys 2004; 31:1512–1520.
28. Tourassi GD, Harrawood B, Floyd CE, Jr. Cross-digitizer robustness of a knowledge-based CAD system for mass detection in screening mammograms. SPIE 2007; 6514:65141Y1–65474Y8.
29. Hajnal JV, Hill DLG, Hawkes DJ. Medical image registration. Boca Raton, FL: CRC Press, 2000.
30. Studholme C, Hill DLG, Hawkes DJ. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognit 1996; 32:71–86.
31. Dekker N, Ploeger LS, van Herk M. Evaluation of cost functions for gray value matching of two-dimensional images in radiotherapy. Med Phys 2003; 30:778–784.
32. Bovik AC, Huang S, Manson DC. The effect of median filtering on edge estimation and detection. IEEE Trans Pattern Anal Machine Intell 1987; 9:181–194.
33. Mayo P, Rodenas F, Verdu G. Comparing methods to denoise mammographic images. IEEE EMBS 2004; 1:247–250.
34. Gabor D. Theory of communication. J Inst Elec Eng 1946; 93:429–459.
35. Zheng Q, Chellappa R. A computational vision approach to image registration. IEEE Trans Imag Processing 1993; 2:311–326.
36. Manjunath BS, Shekhar C, Chellappa R. A new approach to image feature detection with applications. Pattern Recognit 1996; 29:627–640.
37. Liu J, Vemuri BC, Marroquin JL. Local frequency representations for robust multimodal imageregistration. IEEE Trans Med Imag 2002; 21: 462–469.
38. Rubner Y, Tomasi C. Perceptual metrics for image database navigation. Norwell, MA: Springer, 2001.
39. Hubel DH, Weisel TN. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. J Neurophysiol 1965; 28:229–289.
40. Campbell FW, Robson JG. Application of Fourier analysis to the visibility of gratings. J Physiol 1968; 197:551–566.
41. Daugman JG. Complete discrete 2D Gabor transforms by neural networks for image analysis and compression. IEEE Trans ASSP 1998; 36:1169–1179.
42. Chen CC, Chen CC. Gabor transform in texture analysis. SPIE Proc 2003; 2353:237–245.
43. Watson AB, Braddick OJ, Sleigh AC. Detection and recognition of simple spatial forms. In: Braddick OJ, Slade AC, eds. Physical and biological processing of images. Berlin, Germany: Springer-Verlag, 1983.
44. Gonzalez RC, Woods RE, Eddins SL. Digital image processing using MATLAB. Upper Saddle River, NJ: Prentice Hall, 2003.
45. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003; 229:3–8.
46. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996; 201:745–750.
47. Tourassi GD, Jesneck JA, Mazurowski ML, et al. Stacked generalization in computer-assisted decision systems: empirical comparison of data handling schemes. Proc IJCNN 2007; 1343–1347.
48. Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2003.
49. The R Project for Statistical Computing. Available online at: http://www.R-project.org. Accessed January 7, 2008.
50. Tourassi GD, Saunders R, Samei E. Mass detection in full field digital mammograms: validation of an information-theoretic knowledge-based system. RSNA 92nd Scientific Assembly, Chicago, IL, 2006.

Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms

Georgia D. Tourassi[a)] and Brian Harrawood
Digital Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, North Carolina 27705

Swatee Singh, Joseph Y. Lo, and Carey E. Floyd
Digital Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, North Carolina 27705 and Department of Biomedical Engineering, Duke University, Durham, North Carolina 27710

(Received 14 June 2006; revised 3 November 2006; accepted for publication 6 November 2006; published 18 December 2006)

The purpose of this study was to evaluate image similarity measures employed in an information-theoretic computer-assisted detection (IT-CAD) scheme. The scheme was developed for content-based retrieval and detection of masses in screening mammograms. The study is aimed toward an interactive clinical paradigm where physicians query the proposed IT-CAD scheme on mammographic locations that are either visually suspicious or indicated as suspicious by other cuing CAD systems. The IT-CAD scheme provides an evidence-based, second opinion for query mammographic locations using a knowledge database of mass and normal cases. In this study, eight entropy-based similarity measures were compared with respect to retrieval precision and detection accuracy using a database of 1820 mammographic regions of interest. The IT-CAD scheme was then validated on a separate database for false positive reduction of progressively more challenging visual cues generated by an existing, in-house mass detection system. The study showed that the image similarity measures fall into one of two categories; one category is better suited to the retrieval of semantically similar cases while the second is more effective with knowledge-based decisions regarding the presence of a true mass in the query location. In addition, the IT-CAD scheme yielded a substantial reduction in false-positive detections while maintaining high detection rate for malignant masses. © *2007 American Association of Physicists in Medicine.* [DOI: 10.1118/1.2401667]

I. INTRODUCTION

There is conflicting evidence regarding the clinical impact of computer-assisted detection (CAD) systems for the diagnostic interpretation of screening mammograms. For the most part, retrospective studies suggest that CAD technology has a positive impact on early breast cancer detection (e.g., Refs. 1–5). There are, however, several retrospective[6–8] and prospective[9–13] studies that produced contradictory conclusions. Although it is recognized that more prospective studies are needed on the topic, it is well known that radiologists often dismiss correct CAD cues. The radiologists' reluctance to trust CAD is mainly attributed to the higher than desired false positive rate.[11] The above observations are particularly true for the detection of masses, a far more challenging task than the detection of calcifications.

While the true clinical benefit of CAD is still debated,[14] CAD research continues in an effort to improve diagnostic performance and clinical integration.[15] For example, the currently used "black-box" CAD paradigm is rather limited. A CAD system that is more interactive and capable of justifying the visual cues it provides may help radiologists' cognitive process more effectively. Moreover, as clinical image libraries grow rapidly in Radiology, contemporary CAD systems should be able to capitalize on accumulating image data without requiring painstaking retraining or recalibration.

Content-based image retrieval (CBIR) could facilitate the development of a new generation of interactive CAD technology that takes advantage of the vast amounts of digital image data generated in clinical practice. The main objective of CBIR research is to develop a user-friendly framework that allows users to interact with digital image libraries effectively.[16] CBIR has been identified as an important research direction in Radiology to facilitate clinical decision support for medical image interpretation.[17,18]

Shifting the CAD paradigm to incorporate image retrieval capabilities is a challenging proposition. The primary task of CBIR in the clinical arena is to help radiologists retrieve images with similar visual content. Medical image retrieval has traditionally been based on text describing the patient clinical data and medical condition depicted in the patient's imaging studies. These textual descriptors are used as keywords for searching the medical image library. Several researchers have recognized the need for more sophisticated image retrieval methods that capture the visual content of images more effectively than textual descriptors. Consequently, CBIR has evolved toward feature-based similarity assessment. Images are compared and retrieved based on low-level image features that describe the color, shape, texture, and spatial arrangement of important objects (i.e., organs, tumors, etc.) identified in the medical images. Nevertheless, low-level image features are often ineffective in

CBIR of single-modality images due to the subtle differences that exist among same-domain images. This inefficiency is known as the "semantic gap" between image features and the visual and diagnostic content of the images as perceived by the radiologists.[17] Therefore, the challenge in creating clinically effective CBIR-based CAD systems is to develop algorithms that retrieve semantically and perceptually similar images to provide evidence-based decision support.

Working toward this goal, we have previously presented a CBIR-based CAD system for the detection and diagnosis of masses in screening mammograms.[19,20] In contrast to feature-based CBIR algorithms in mammography,[21–27] our system relies on information theoretic principles to assess image similarity. Specifically, the system uses the popular concept of mutual information (MI) to measure the similarity between a query image and those stored in the knowledge database. MI-based similarity assessment relies completely on the statistical properties of the image histograms eliminating the image preprocessing, segmentation, and feature extraction steps. Furthermore, information theoretic similarity measures have the advantage of making no assumptions on the underlying image distributions. Our CAD system was evaluated initially as a knowledge-based system for the discrimination of masses from normal breast parenchyma[19] and for the diagnostic characterization of masses using relevance feedback techniques.[20]

Since similarity assessment is the most important component in CBIR,[28,29] the purpose of this study was to explore several entropy-based similarity measures for region-based analysis of mammograms. Specifically, we present a comparative study using the information-theoretic computer-assisted detection (IT-CAD) scheme for three clinically oriented tasks. First, an experiment was performed to determine which similarity measure helps the IT-CAD scheme retrieve semantically relevant mammographic regions with the highest precision. A second experiment was performed to determine which measure helps the IT-CAD scheme discriminate between mass and normal mammographic regions with the highest accuracy. Finally, a third experiment was performed to validate the conclusions of Experiments 1 and 2 using IT-CAD for evidence-based, false positive reduction of progressively more challenging visual cues produced by an existing second-reader CAD system.

II. MATERIALS AND METHODS

A. Information-theoretic similarity measures

Information-theoretic (dis)similarity measures are based on the concept of entropy.[30] The most commonly used entropy definition is the Shannon entropy (H):

$$H = -\sum_{x} p(x)\log_2[p(x)], \tag{1}$$

where $p(x)$ is the probability that an image pixel will have the intensity value x. The image probability $p(x)$ is typically estimated from the image histogram, commonly using the convention $0 \log 0 = 0$. Entropy is considered a measure of the uncertainty or complexity in an image. The image complexity (or uncertainty) is captured by the dispersion of the probability distribution of the image intensity levels. Images with uniform pixel intensity distributions have high dispersion and therefore higher entropy. In contrast, images with intensity distributions that depict a few large peaks have lower dispersion and thus lower entropy.

Generally, information-theoretic similarity measures compare the histograms of two images X and Y. The comparison may focus only on corresponding histogram bins (i.e., bin-by-bin measures) or it may incorporate information for non-corresponding bins (i.e., cross-bin measures). This study investigates eight information-theoretic (IT) (dis)similarity measures that have been successfully applied in other areas of medical imaging such as image registration, segmentation, and feature-based image retrieval. Four of them are cross-bin measures: (i) joint entropy, (ii) conditional entropy, (iii) mutual information, and (iv) normalized mutual information. The remaining four IT measures are typical examples of bin-by-bin measures: (i) average Kullback-Leibler divergence, (ii) maximum Kullback-Leibler divergence, (iii) Jensen divergence and, (iv) arithmetic-geometric mean divergence. The following is a brief description of each measure.

1. Joint entropy

Joint entropy (JOINT_H) is the entropy of the joint histogram of two images X and Y.

$$JOINT\_H = H(X,Y) = -\sum_{x}\sum_{y} p_{XY}(x,y)\log[p_{XY}(x,y)]. \tag{2}$$

If two images are completely unrelated, their joint entropy is equal to the sum of their individual entropies. On the other hand, the more similar two images are, the lower their joint entropy is compared to the sum of the individual entropies. Consequently, the joint entropy is a distance measure rather than a similarity measure. Two images with lower joint entropy are considered more similar (i.e., more relevant) than two images with higher joint entropy.

2. Conditional entropy

The conditional entropy $H(X|Y)$ of two images X and Y measures how much entropy (or uncertainty) is remaining regarding image X (i.e., the query image) when we have learned the truth regarding image Y (i.e., an image in the knowledge database). Similarly to joint entropy, conditional entropy is also a dissimilarity measure. Therefore, if two images are relevant, then the conditional entropy (or uncertainty) of the query image given the known image should be low. However, in contrast to joint entropy, conditional entropy is not symmetric. In other words, $H(X|Y) \neq H(Y|X)$. The conditional (COND_H) and joint entropy of two images X and Y are related as follows:

$$COND\_H = H(X|Y) = H(X,Y) - H(Y) \tag{3}$$

3. Mutual information

Mutual information (MI) is the most popular IT similarity measure, particularly for image registration.[31–33] MI is similar to joint entropy but it also takes into account the individual image entropies.

$$MI(X,Y) = H(X) + H(Y) - H(X,Y)$$
$$= \sum_x \sum_y P_{XY}(x,y)\log_2 \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}. \quad (4)$$

MI is a measure of general statistical dependence (i.e., shared information) between two images. It measures the amount of uncertainty reduction about one image given the information we have about the other image. MI is a true similarity measure. The more similar X and Y are, the higher their MI. Furthermore, MI is considered a generalized extension of the correlation coefficient because it does not make linear assumptions regarding the relationship between the two images' pixel values.[34]

4. Normalized mutual information

Normalized mutual information (NMI) is a normalized version of MI ensuring that the similarity measure is bounded between 0 and 1. Previous studies in image registration have shown that NMI is often more successful and robust than MI (Ref. 31).

$$NMI(X,Y) = \frac{H(X) + H(Y)}{H(X,Y)}. \quad (5)$$

5. Relative entropy

Relative entropy or Kullback-Leibler (KL) divergence is a distance measure between two probability distributions $p(x)$ and $q(x)$. In the scope of this study, $p(x)$ and $q(x)$ are the probability distributions of the stored image $p(x)$ and the query image $q(x)$, respectively. The relative entropy is defined as follows:

$$D(q \parallel p) = \sum_x q(x)\log[q(x)/p(x)]. \quad (6)$$

Relative entropy is typically used in coding theory and it measures how inefficient on average it would be to use the histogram of one image to code another. Generally, the higher the relative entropy is, the more dissimilar the two images are. Similarly to conditional entropy, KL divergence is not a true distance measure because it is not symmetric (i.e., $D(q\|p) \neq D(p\|q)$). Consequently, different transformations have been utilized in CBIR to provide a symmetric KL divergence measure (SKL).[35] In this study, we have explored two such transformations: (i) the average KL divergence

$$SKL\_1 = \frac{D(q \parallel p) + D(p \parallel q)}{2} \quad (7)$$

and (ii) the maximum KL divergence

$$SKL\_2 = \max[D(q \parallel p), D(p \parallel q)]. \quad (8)$$

SKL is a non-negative distance metric that is equal to 0 when the two probability distributions are identical.

6. Jensen divergence

Some studies have indicated that KL divergence $D(p\|q)$ is not numerically stable and is often sensitive to histogram binning.[36] Consequently, another divergence measure has been proposed as a more stable alternative. The Jensen divergence (JD) is an empirical modification of the KL divergence that is symmetric and more robust with respect to noise and histogram binning[36]

$$JD(p,q) = \sum_x \left(q(x)\log \frac{2q(x)}{p(x) + q(x)} + p(x)\log \frac{2p(x)}{p(x) + q(x)} \right). \quad (9)$$

The Jensen divergence has values bounded between 0 and 2.

7. Arithmetic-geometric mean divergence

Finally, the last similarity measure explored was the arithmetic-geometric mean (AGM) divergence. This measure is essentially the KL divergence between the arithmetic and geometric means of the two image distributions $p(x)$ and $q(x)$

$$AGM(p,q) = \sum_x \frac{p(x) + q(x)}{2} \log \frac{p(x) + q(x)}{2\sqrt{p(x)q(x)}}. \quad (10)$$

All above IT measures require estimation of the marginal probability distribution of the individual images. In addition, some measures (i.e., JOINT_H, COND_H, MI, NMI) require estimation of the joint probability distribution of the two images as well. Consistent with our earlier study[19] and for reasons of computational efficiency, we applied the histogram approach[33] to approximate the marginal and joint probability distributions functions. The number of histogram bins for histogram approximation was selected empirically. We varied the number of histogram bins (i.e., 4, 8, 16, 32, 64, 128) and repeated the experiments with respect to all similarity measures. As expected, the number of histogram bins affected the observed results. For example, using only four bins produced consistently inferior results across all similarity measures. The differences among the results observed for the remaining values of the histogram bin parameter were not statistically significant. Overall, 64 bins were sufficient for histogram approximation across all similarity measures and clinical tasks. For each ROI, the mean μ and standard deviation σ of the ROI pixel values were calculated. Then, the interval $[\mu-2\sigma, \mu+2\sigma]$ was divided into 64 equal-sized bins. Pixel values falling outside the predetermined $[\mu-2\sigma, \mu+2\sigma]$ interval were assigned to the outermost bins when calculating the histograms. The above rules were followed consistently for all images, similarity measures, and experiments.

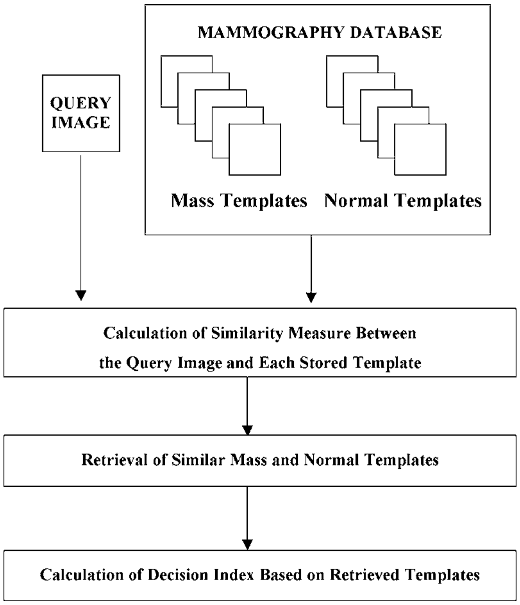

QUERY IMAGE
MAMMOGRAPHY DATABASE
Mass Templates
Normal Templates
Calculation of Similarity Measure Between the Query Image and Each Stored Template
Retrieval of Similar Mass and Normal Templates
Calculation of Decision Index Based on Retrieved Templates

FIG. 1. Schematic representation of the IT-CAD content-based retrieval and detection scheme for mammographic masses.

B. Overview of the information-theoretic CAD system

Figure 1 shows a schematic representation of the image retrieval scheme with the proposed information theoretic framework for mass detection. The scheme is designed to provide region-based evaluation of mammograms for a targeted, evidence-based analysis of suspicious mammographic locations.

Initially, a query mammographic location is presented to the IT-CAD system. The system extracts a fixed size mammographic region around the specific location. The query region of interest (ROI) is compared to a knowledge database of ROIs with known ground truth. Similar cases are retrieved from the knowledge database. A decision is formulated regarding the query region using the retrieved similar cases.

There are two critical components in the IT-CAD scheme: (i) the similarity measure, and (ii) the knowledge database. Since the clinical focus of the IT-CAD scheme is mass detection, it is reasonable to expect that the knowledge database should contain a rich collection of mammographic ROIs that depict biopsy-proven masses. Although the above requirement is critical, the knowledge database also includes a diverse set of ROIs that depict normal breast parenchyma. Because the similarity measure is calculated using the full ROI, it is possible that two ROIs may result in high similarity mainly due to parenchymal background similarities rather than the potential abnormalities they contain. Consequently, the information theoretic CAD approach decides based on both similar mass and normal cases that are stored in its knowledge database. Specifically, the IT-CAD decision index $D(Q)$ is calculated as follows:

$$D(Q) = \frac{1}{k}\sum_{j=1}^{k} SM(Q,M_j) - \frac{1}{k}\sum_{j=1}^{k} SM(Q,N_j), \tag{11}$$

where Q is the query mammographic region, SM stands for similarity measure, and M_j and N_j are known mass and normal cases that are retrieved from the knowledge database as most similar to the query. Note that if the query region depicts a mass, it is expected that the calculated decision index should be higher than if it contains normal parenchyma. The second term in Eq. (11) is a correction term so that high values of $D(Q)$ are less likely to be the result of matching backgrounds than matching potential abnormalities.

Although our previous studies have shown promising results using mutual information as the similarity measure (SM), this study explores several other information-theoretic (dis)similarity measures that share the same featureless simplicity and computational efficiency with MI. Note that in Eq. (11), SM denotes a similarity measure. The proposed dissimilarity measures (i.e., joint entropy, conditional entropy, KL divergence, Jensen divergence, and geometric/arithmetic mean) can be easily converted into similarity measures by taking their negative or inverse value. For this study, we applied the negative transformation.

C. Data collection and study design

The study was based on 512×512 pixel ROIs extracted from mammograms for the Digital Database of Screening Mammography (DDSM).[37] The mammograms are 12 bit images digitized using the Lumisys scanner at 50 μm per pixel. No image preprocessing (i.e., segmentation, filtering, normalization, etc.) was performed on the mammograms or the extracted ROIs.

We created two different ROI databases based on the DDSM/Lumisys mammograms. Database 1 contained 1820 ROIs. Of those, 901 ROIs depicted a biopsy-proven mass (489 malignant and 412 benign). The ROIs were centered around the physician's annotation provided in the DDSM truth files. The remaining ROIs were extracted from 62 normal mammograms (two ROIs per breast, per view) for a total of 8×62=496 normal ROIs. The location of the normal ROIs was selected randomly within the breast. There was no overlap between the ROIs extracted from the same image. To keep the database evenly balanced between normal and abnormal ROIs, an additional 424 ROIs were extracted from abnormal DDSM/Lumisys cases, but only from breasts that did not contain any physician annotations in either mammographic view. The selection of these cases was random. Therefore, the final database contained 919 ROIs that were deemed normal.

Database 2 contained ROIs extracted from 100 DDSM cases completely different from those used to create Database 1. This second database was selected to represent a balanced mix of abnormal and normal cases from all available DDSM/Lumisys volumes. Note that the DDSM volumes correspond to patient data acquired at different geographic locations. By creating a balanced mix of cases we tried to minimize potential discrepancies due to patient population differences. Furthermore, within each volume an equal number of cases were selected for each mammographic density. Of the 100 DDSM cases in Database 2, 40 cases contained malignant masses, 40 cases contained benign masses, and the remaining 20 cases were considered normal. In DDSM a screening mammogram is considered normal if it does not require any further "follow-up," it does not contain any annotated abnormalities, and the patient has a normal screening exam at least four years later.

Database 2 was processed using a previously presented, in-house CAD system for mass detection.[38,39] The system was used to locate suspicious locations within the images. The CAD system is a multi-stage algorithm consisting of a typical sequence of steps: (i) image filtration using a difference of Gaussians filter,[40,41] (ii) initial localization of suspicious regions detected at high sensitivity using a progressive gray level thresholding procedure, (iii) feature extraction and selection, and (iv) feature-based classification using Fisher's linear discriminant for false positive reduction of the initial suspicious regions. The prescreening, in-house CAD system was initially trained and optimized on a separate set of DDSM cases, completely different from Database 2. After training and optimization, the system was applied "as is" on Database 2.

Specifically, the in-house, mass detection system was applied on the craniocaudal (CC) views of the 80 cases in Database 2 that contained the annotated masses. For the 20 normal cases in Database 2, only one, randomly selected CC view (left or right breast) was analyzed. Therefore, 100 independent images were analyzed. The automated screening process resulted in 399 false positive (FP) detections (approximately 4 FPs/image). In addition, depending on the definition of true positive detection,[42] the system also detected 84%–92% of the true masses. However, because our main focus is on reducing further the false positive detections, we combined the 399 FPs with all true masses annotated in the 100 images anticipating future sensitivity improvement of our prescreening algorithm. In total, there were 483 mammographic regions in Database 2; 44 depicting a malignant mass, 40 depicting a benign mass, and 399 depicting suspicious looking yet normal breast parenchyma.

Database 1 was used in a leave-one-out manner to assess how the various image similarity measures impact the retrieval precision (Experiment 1) and diagnostic accuracy (Experiment 2) of our IT-CAD scheme. The leave-one-out sampling scheme was implemented on a per case basis as follows. Each ROI in Database 1 was excluded once to serve as the query. Of the remaining 1819 ROIs, the ones extracted from DDSM cases different than the query's served as the knowledge database of the IT-CAD scheme. The same process was repeated until each ROI served as a query.

Experiment 3 aimed to validate the conclusions drawn from experiments 1 and 2 for the clinical task of reducing the false positive detections of prescreening CAD systems. Both Databases 1 and 2 were used for this third experiment. Specifically, the ROIs in Database 2 served as queries for testing the IT-CAD system while Database 1 served as the knowledge database.

D. Performance evaluation

Two different performance indices were employed in this study depending on the operating mode of the IT-CAD system (retrieval engine vs. detection aid). When the system was tested as a retrieval engine, its retrieval capabilities were assessed using precision as the selected performance index. Given a query image, precision (P) is the number of relevant retrieved images (R) divided by the total number of retrieved images (K)

$$Precision(P) = \frac{Number\ of\ relevant\ retrieved\ images\ (R)}{Total\ number\ of\ retrieved\ images\ (K)}. \tag{12}$$

Retrieval precision in CBIR is analogous to positive predictive value in decision analysis. There are two ways to define relevance in image retrieval; visual or semantic. For this application, we focus on semantic relevance. A retrieved image is considered to be relevant if it belongs to the same class (mass or nonmass) as the query image. Since retrieval precision is dependent on the query, a CBIR system's precision is typically reported averaged across all queries. According to Eq. (12), retrieval precision is also dependent on the number of retrieved images (K) and it is typically plotted as a function of K. In this study, we focus only on the top 1, 5, and 10 retrievals and evaluate the eight similarity measures with respect to these top retrieved cases. We limited the precision analysis to $K \leqslant 10$ for practical reasons. In an interactive CAD system, it is impractical to present radiologists with more than the top ten most similar cases for visual evaluation.

Receiver operating characteristic (ROC) analysis[43] was employed to assess the performance of the IT-CAD system as a mass detection aid. The decision index calculated based on Eq. (11) was used as the decision variable for ROC analysis. Since the decision index is dependent on the number of closest mass and normal retrievals (k), ROC analysis was performed for a wide range of k values. The ROC analysis was performed using the ROCKIT software developed by Charles Metz at the University of Chicago.

III. RESULTS

A. Experiment 1: Retrieval precision

The average retrieval precision achieved by each similarity measure at the top $K=1$, 5, and 10 retrievals was calculated for all queries and separately for each subgroup of queries (i.e., malignant masses, benign masses, and normals). Subgroup analysis was performed to identify possible discrepancies depending on the true class of each query ROI. Overall there were subtle changes as the number of retrievals increased from $K=1$ to $K=10$. Thus, Fig. 2 shows results for the top $K=1$ and $K=5$ retrievals only [Figs. 2(a) and 2(b), respectively].

Figure 2 shows that the overall retrieval precision $P(K)$ achieved by the eight similarity measures appears to be within the range of 47%–65%. Bin-by-bin measures demonstrated overall higher average retrieval precision compared to

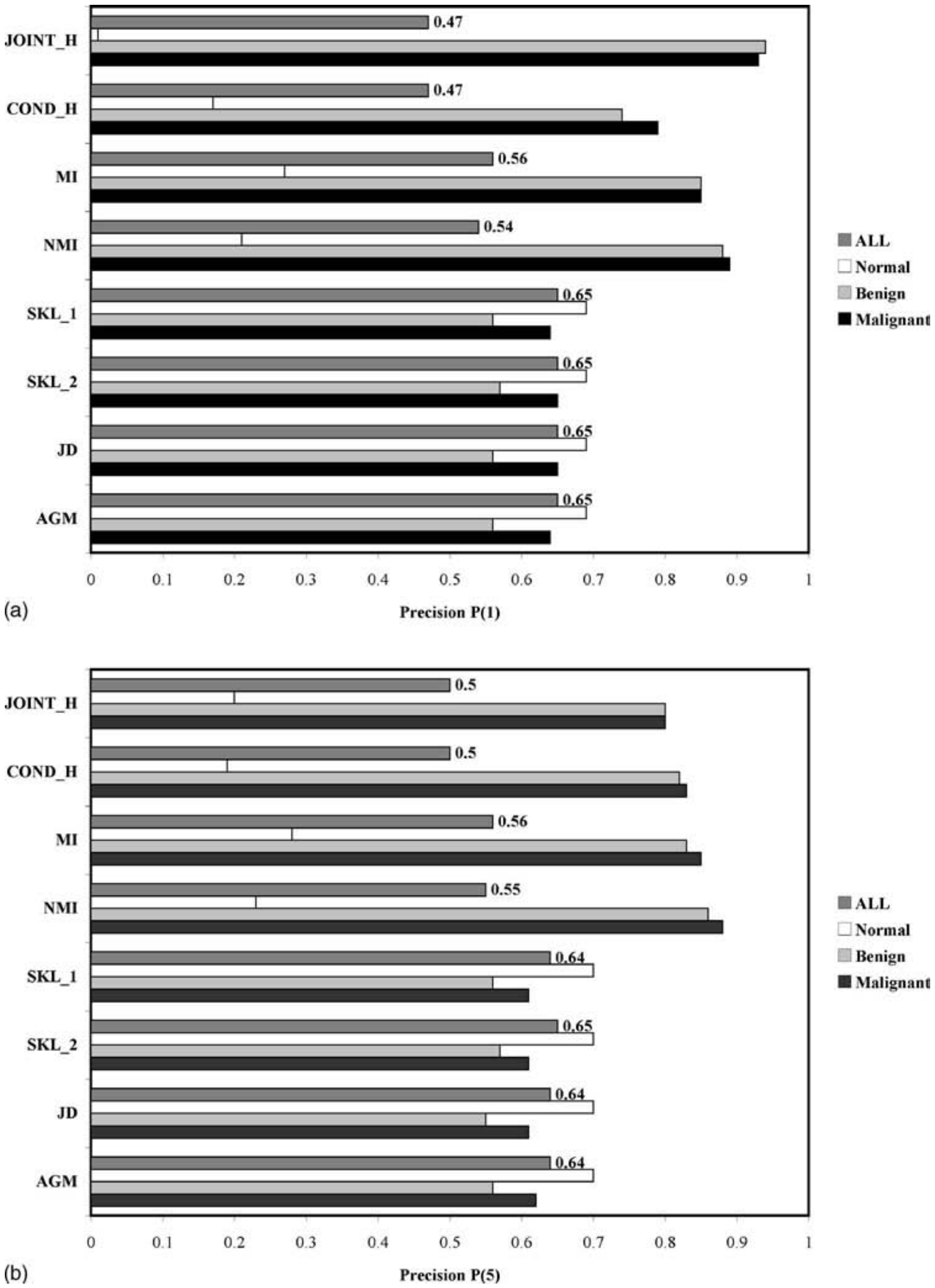


FIG. 2. Average retrieval precision $P(K)$ for the (a) top $K=1$ and (b) $K=5$ retrievals for all similarity measures. Precision is shown overall and for each subgroup of query cases separately. (JOINT_H: joint entropy, COND_H: conditional entropy, MI: mutual information, NMI: normalized mutual information, SKL: symmetric Kullback-Leibler divergence, SKL_MAX: maximum Kullback-Leibler divergence, JD: Jensen divergence, AGM: arithmetic-geometric mean divergence).

the cross-bin measures at all three retrieval levels. In addition, there were dramatic differences depending on the type of query. Four similarity measures (i.e., joint entropy, conditional entropy, mutual information, and normalized mutual information) achieved significantly higher precision for mass queries rather than normal queries. In contrast, average precision was far more robust between mass and normal queries for the remaining similarity measures (i.e., symmetric Kullback-Leibler divergence, maximum Kullback-Leibler divergence, Jensen divergence, and arithmetic-geometric mean divergence). However, the average retrieval precision was consistently higher for normal queries than masses for the second group of similarity measures. It is notable that the average retrieval precision for malignant masses was consistently higher than that for benign masses for all similarity measures (with the only exception for the joint entropy measure at the top $K=1$ retrieval).

Note that since the normal and mass ROIs are almost evenly balanced in Database 1, there is a 50% chance to randomly retrieve a mass or normal template from the knowledge database. The Wilcoxon signed rank test was performed to determine if the average precision was significantly higher than the expected ~50% precision value due to the inherent prevalence of each subgroup (i.e., 49.5% for mass and 51.5% for normal ROIs) in the database. For all K values, all similarity measures, and all subgroups of query cases the observed precision was statistically significantly different (p value<0.0001) than the expected ~50% aver-

TABLE I. ROC area index A_z (±0.01) achieved by the IT-CAD scheme depending on the similarity measure and the number of the best-matched k mass and normal templates considered in decision making.

k	JOINT_H	COND_H	MI	NMI	SKL_1	SKL_2	AGM	JD
1	0.47	0.48	0.71	0.71	0.68	0.70	0.67	0.70
5	0.48	0.50	0.78	0.79	0.73	0.75	0.74	0.75
10	0.51	0.51	0.81	0.81	0.76	0.77	0.76	0.76
50	0.59	0.57	0.87	0.85	0.77	0.77	0.77	0.77
100	0.67	0.70	0.87	0.85	0.76	0.76	0.75	0.76
300	0.85	0.85	0.86	0.86	0.73	0.73	0.73	0.73
500	0.86	0.86	0.86	0.87	0.63	0.64	0.62	0.65
700	0.87	0.87	0.87	0.87	0.60	0.61	0.60	0.62
ALL	0.86	0.86	0.87	0.87	0.59	0.59	0.58	0.63

age precision if retrieval were purely random. This result was consistent for subgroups and similarity measures where the achieved precision was significantly inferior to that expected with random retrieval (e.g., 20% average precision $P(5)$ for normal ROIs using joint entropy as the similarity measure).

The signed rank test with Bonferroni correction for multiple comparisons was also performed to test for significant differences in average retrieval precision among the different similarity measures. The analysis was performed for each K value ($K=1,5,10$) and each query subgroup (malignant, benign, normal) separately at the 95% confidence level. The consistent trend among the results was that the four dissimilarity measures SKL_1, SKL_2, JD, AGM provide very similar average retrieval precision for all query groups.

On the other hand, the remaining four similarity measures (JOINT_H, COND_H, MI, NMI) provide significantly different precision performance compared to the first group across all subgroups and retrieval levels ($K=1,5,10$). Indeed, nonparametric correlation analysis confirmed that (SKL_1, SKL_2, JD, AGM) and (JOINT_H, COND_H, MI, NMI) represent two distinct groups of measures. The similarity measures of the first group resulted in highly correlated precision performance for all query groups ($0.87 \leqslant \rho \leqslant 0.98$). However, the similarity measures of the second group resulted in significantly less correlated precision performance ($-0.14 \leqslant \rho \leqslant 0.62$) with the exception of COND_H and MI ($0.65 \leqslant \rho \leqslant 0.90$ depending on the query group and number of top retrievals). Surprisingly, the mutual information and normalized mutual information measures resulted in lower correlation ($0.56 \leqslant \rho \leqslant 0.84$ depending on the query group and number of top retrievals). It is noted that the differences in precision between MI, NMI, COND_H, and JOINT_H were often significant for both mass and normal queries at the various retrieval levels. The above statistical analysis was performed using the JMP Statistical Software Version 5.1 available from SAS, Cary, NC.

B. Experiment 2: Detection accuracy

The (dis)similarity measures were subsequently used in the IT-CAD system for the discrimination of mass from normal ROIs according to the decision variable described in Eq. (11). In contrast to retrieval precision, the decision variable ignores the rank order of the retrieved cases but it takes into consideration the actual value of the similarity measure under consideration.

Table I shows the corresponding ROC areas achieved for each similarity measure based on the number k of the closest mass and normal templates retrieved from the knowledge database. Results are shown for several k values to highlight the general trends. For example, when $k=1$, the IT-CAD system is asked to make a decision using the one mass and one normal templates retrieved from the database as most similar to the query. In contrast, if $k=\text{ALL}$, the IT-CAD system is asked to make a decision using the whole knowledge database.

Using the mutual information, normalized mutual information, conditional entropy, and joint entropy as the similarity measure, the IT-CAD system achieved its highest ROC performance ($A_z=0.87\pm0.01$). Although not shown in Table I, the IT-CAD performed significantly better for the detection of malignant ($A_z=0.89\pm0.01$) than benign masses ($A_z=0.84\pm0.01$). The number of top mass and normal templates required for optimized performance depended on the similarity measure. Using mutual information, the system achieved its highest performance using as few as the top matched 50 mass and normal templates. Conditional entropy, normalized mutual information, and joint entropy required substantially more matched templates. The best ROC area index achieved by the IT-CAD scheme was significantly lower when using the Kullback-Leibler, Jensen, and arithmetic-geometric mean divergence measures ($A_z=0.77\pm0.01$). This performance was optimized with approximately 50 best matched mass and normal templates and deteriorated substantially as more inferior matches were included in the decision making process.

Since emphasis is typically place on operating at a high sensitivity level for breast cancer detection tasks, the impact of the eight similarity measures was also evaluated with respect to the partial ROC area index ${}_{0.90}A_z$. The overall trends remained the same. Specifically, MI, NMI, JOINT_H, and COND_H achieved significantly higher performance for malignant masses (${}_{0.90}A_z=0.57\pm0.03$) than the remaining measures (${}_{0.90}A_z=0.31\pm0.03$).

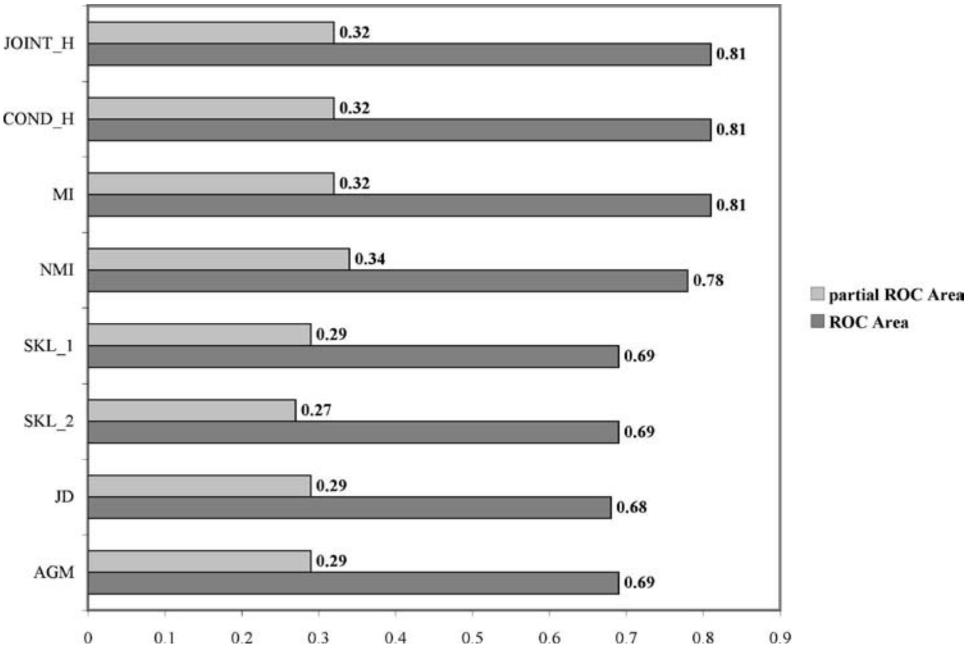


FIG. 3. Overall ROC area (±0.02) and partial ROC area (±0.04) indices achieved by the IT-CAD scheme for discrimination of true masses from suspicious yet normal ROIs depending on the image similarity measure.

C. Experiment 3: IT-CAD for false positive reduction

Finally, the IT-CAD scheme was validated for discriminating true masses from false positive mammographic regions. Figure 3 shows the overall ROC and partial ROC area indices achieved depending on the similarity measure. As the figure shows, the best performance of the IT-CAD scheme (A_z=0.81±0.02) is significantly lower on Database 2 than what was previously observed on Database 1 (A_z =0.87±0.01). The performance deterioration was expected because Database 2 represents a far more challenging detection task (mass vs. suspicious-looking normal ROIs) than Database 1 (mass vs. randomly chosen normal ROIs). However, the same overall trends prevailed. The same two groups of similarity measures emerged with distinctly different detection performance.

The IT-CAD detection performance on Database 2 was analyzed in more detail with NMI as the similarity measure. NMI is a more attractive choice than MI due to its bounded nature. It always ranges between 0 and 1 regardless of any possible preprocessing done on the ROIs. Three operating decision thresholds were selected using the partial ROC curves acquired in Database 1. The thresholds corresponded to three clinically relevant operating points: (a) 95% sensitivity, (b) 90% sensitivity, and (c) 85% sensitivity for malignant masses. Note that the decision thresholds for these three operating points were determined using Database 1 exclusively. Database 2 was used purely for testing.

Overall, the IT-CAD scheme had very robust detection performance in Database 2 when operating with the above decision thresholds. Operating at the desired 95% sensitivity decision threshold, the IT-CAD system detected 42/44 malignant masses present in Database 2 (95.7% sensitivity). At the 90% and 85% sensitivity operating thresholds, the scheme detected 40/44 malignant masses (90.9% sensitivity).

The contribution of the IT-CAD system for false positive reduction rate was also assessed at the same three operating points. Table II shows the false positive reduction rate for all false positives and for progressively more challenging ones; those remaining when the system operates at 3 FPs/image, 2 FPs/image, 1 FP/image, and 0.4 FP/image, respectively (at the expense of lower mass detection rate obviously). The IT-CAD scheme can effectively reduce about 50% of the false positive cues while detecting 90% of the malignant masses. Although the impact of the IT-CAD scheme deteriorates as the false positive cues become progressively more challenging, the scheme can still eliminate up to 17.5% of the false positive cues generated by the prescreening system that operates at a low 0.4 FP/image (while detecting 85% of malignant masses).

The results regarding retrieval precision in Database 2 were also consistent with what was observed in Database 1. The bin-by-bin similarity measures (SKL_1, SKL_2, JD, AGM) provided far more robust retrieval precision among all queries than the cross-bin similarity measures (JOINT_H, COND_H, MI, NMI). It is noted however, that the average retrieval precision for the false positive ROIs was consistently lower than that achieved for randomly chosen normal ROIs in Database 1 using the bin-by-bin similarity measures (0.60 vs 0.69).

IV. DISCUSSION

Assessment of image similarity is a critical step for the retrieval and diagnostic interpretation of medical images based on their content. The task involves two important decisions: (i) how to represent the image, and (ii) how to choose the most effective similarity measure for the specific image representation space and the particular medical task at hand. Typically, these decisions are made empirically using a

TABLE II. False positive reduction rate achieved by the IT-CAD scheme stratified according to the difficulty level of the cases. Results are shown for three malignant mass sensitivity [true positive fraction (TPF)] operating points. The IT-CAD scheme employed normalized mutual information as the similarity measure.

DIFFICULTY LEVEL OF FP CUES (No. of FPs/image)	% FP REDUCTION (Remaining average No. of FPs/image)		
	TPF=95%	TPF=90%	TPF=85%
ALL FPs (4 FPs/img)	29.8% (2.80)	45.9% (2.16)	52.0% (1.91)
75% most challenging FPs (3 FPs/img)	21.0% (2.37)	38.0% (1.86)	45.7% (1.63)
50% most challenging FPs (2 FPs/img)	14.0% (1.72)	29.5% (1.41)	36.0% (1.28)
25% most challenging FPs (1 FPs/img)	10% (0.90)	18.0% (0.82)	25.0% (0.75)
10% most challenging FPs (0.4 FPs/img)	5% (0.38)	7.5% (0.37)	17.5% (0.33)

labeled database. The size and comprehensiveness of the database usually determine how well the decisions generalize to new databases.

In the present study we investigated the retrieval performance and mass detection accuracy of eight information-theoretic (IT) image similarity measures for region-based analysis of mammograms. In contrast to feature-based similarity assessment techniques, the IT measures operate with image histograms without requiring image feature extraction. Thus, the image content is represented in terms of pixel intensity histograms. The IT similarity measures essentially compare the region-based histograms of two mammograms to determine how relevant they are. Specifically, this study focused on two groups of information theoretic measures: (i) bin-by-bin measures that compare only the contents of corresponding histogram bins (i.e., average KL divergence, maximum KL divergence, Jensen divergence, arithmetic-geometric mean divergence) and (ii) cross-bin measures (i.e., joint entropy, conditional entropy, mutual information, normalized mutual information) that incorporate the comparisons of the contents of noncorresponding bins.

The proposed image similarity measures were evaluated in the context of an interactive CAD system that is designed to provide evidence-based decisions regarding the presence of a malignant mass in mammographic locations that serve as queries for the system. These measures were evaluated in two different capacities: (i) for retrieval of diagnostically similar cases and for (ii) knowledge-based mass detection. Experiments were performed using two independent databases. The first database contained mammographic regions that depicted either a mass or normal breast parenchyma. This database was used for empirical comparison of the similarity measures based on a leave-one-case-out sampling scheme. The main conclusions drawn from using Database 1 were further validated on Database 2. The second database served as a clinically more challenging test bed because the nonmass mammographic regions it contained were already cued as highly suspicious for containing a mass by an inhouse CAD system. Therefore, the additional validation experiment aimed to evaluate to what extent the information-theoretic CAD analysis could improve upon the performance of existing CAD technology by providing evidence-based analysis of suspicious regions.

Our study clearly demonstrated two strong trends. First, bin-by-bin measures based only on the distance of the marginal histograms were more successful at achieving higher and more balanced average retrieval precision of cases with similar semantic content. High precision in the first few retrievals is critical for content-based image retrieval systems designed to display the top matches for visual evaluation by the CBIR user. On the other hand, cross-bin similarity measures that incorporate the joint histogram information were more successful for knowledge-based discrimination of masses from normal mammographic regions.

Based on the above observations, it seems reasonable to consider the ratio of retrieved masses over the total number of retrieved cases as a potential decision variable for the IT-CAD system. Basically, when a query case is presented for evaluation, the IT-CAD scheme retrieves the top K most similar cases. The prevalence of masses in the top retrievals is treated as a predictive variable for the presence of mass in the query image. This predictive variable is in essence similar to the odds ratio. If the query depicts a mass, then the above prevalence should be larger than if the query depicts normal breast parenchyma. Although not reported in this study, we explored this possibility with all similarity measures. As expected, the bin-by-bin similarity measures helped the IT-CAD scheme achieve a higher ROC area index than the cross-bin measures (0.74 ± 0.01 vs. 0.69 ± 0.01) for a low number of retrievals ($K < 30$). As more retrieved cases were considered, the ROC performance evened out between both groups of similarity measures. However, the ROC area index never exceeded the one achieved using the knowledge-based decision index [Eq. (11)] proposed in our study.

Finally, our study showed that the IT-CAD system can be effectively utilized as an add-on to existing detection schemes for false-positive reduction. Since the information-theoretic system follows a featureless-based image analysis, it appears to complement feature-based CAD schemes. Specifically, the IT-CAD system safely eliminated up to 17.5% of the most challenging false positive cues (those generated by prescreening the mammograms with a system that generates 0.4 FP/image) while still detecting 85% of the malignant masses. On a side note, the detection rate achieved for the benign masses was 90% (36/40).

To summarize, our study represents a critical step toward an interactive CAD system able to operate as an effective content-based image retrieval and knowledge-based mass detection system. The comparative analysis demonstrated that the choice of the similarity measure depends on the clinical task (retrieval vs. detection). No particular similarity measure emerges as the optimal choice for both tasks. While MI and NMI appear to be excellent choices for knowledge-based mass detection, they fail to provide robust retrieval precision across the two query classes for the top retrievals. Therefore, these measures are not suitable for CAD users who would like to view the top most relevant matches. In contrast, the bin-by-bin similarity measures such as Jensen divergence and Kullback-Leibler divergence achieved overall higher and more robust retrieval precision. However, these measures failed to reach the detection accuracy achieved by the cross-bin similarity measures. An interesting observation was that regarding retrieval precision, the cross-bin measures resulted in substantially lower pairwise correlation than the bin-by-bin measures. This finding suggests that MI, NMI, JOINT_H, and COND_H are good candidates for a possible fusion retrieval strategy. In fact, the newest trends in content-based image retrieval suggest that composite similarity measures may be more effective than single similarity measures. Our study certainly points toward that direction. For example, a composite strategy where a bin-by-bin similarity measure is used for initial retrieval of semantically similar cases while a cross-bin similarity measure is subsequently used for knowledge-based analysis of the retrieved cases appears to be a promising strategy to achieve simultaneously high retrieval precision and detection accuracy. We are currently investigating this idea.

One of the limitations of the present study design is that it assessed retrieval precision based on semantic, not visual content. This aspect is important for interactive CBIR-based CAD systems. It is possible that cross-bin measures may be more effective at capturing visual content than bin-by-bin measures. We plan to investigate this possibility in the future. In addition, we will investigate how beneficial the system is with mammographic regions that raise visual suspicion. The present study focused only on image locations marked as suspicious by another CAD algorithm. Analyzing image locations that are marked as suspicious by radiologists will determine the role of the IT-CAD system for reducing the interpretation, not perceptual, error associated with the diagnostic interpretation of mammograms.

From a theoretical point of view, the inherent limitation of the information-theoretic measures evaluated in this study is that they focus on the global histograms representation. Therefore, the localized spatial relationships among the image pixels are lost. This limitation has been addressed before in the context of image registration. It has been proposed that taking into account the neighborhood of regions of corresponding image pixels may be a more effective strategy.[44,45] The computational complexity and less dramatic than anticipated improvements of this approach have led researchers to seek simpler surrogate approaches. For example, Pluim, Maintz, and Ueirgever proposed multiplying the mutual information with an additional term that incorporates the local gradients of the two images in comparison.[46] Certainly advances made toward this direction in image registration may have significant implications for our CAD application as well.

In conclusion, this study represents a comprehensive step toward a framework of entropy-based, image similarity assessment for retrieval of diagnostically relevant images to support interactive, evidence-based diagnostic interpretation of mammograms.

ACKNOWLEDGMENTS

This work was supported by Grant No. R01 CA101911 from the National Cancer Institute and by Grant No. W81XWH-05-1-0293 from the Army Breast Cancer Research Program. We would like to thank Dr. Alan Baydush and Dr. David Catarious for providing guidance in the application of the prescreening CAD system to generate the suspicious mammographic regions.

[a] Electronic mail: georgia.tourassi@duke.edu

[1] L. J. W. Burhenne *et al.*, “Potential contribution of computer-aided detection to the sensitivity of screening mammography,” Radiology **215**, 554–562 (2000).

[2] R. L. Birdwell, D. M. Ikeda, K. F. O’Shaughnessy, and E. A. Sickles, “Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection,” Radiology **219**, 192–202 (2001).

[3] R. F. Brem *et al.*, “Improvement in sensitivity of screening mammography with computer-aided detection: A multiinstitutional trial,” AJR, Am. J. Roentgenol. **181**(3), 687–693 (2003).

[4] R. F. Brem and J. M. Schoonjans, “Radiologist detection of microcalcifications with and without computer-aided detection: A comparative study,” Clin. Radiol. **56**(2), 150–154 (2001).

[5] M. A. Helvie *et al.*, “Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection: Pilot clinical trial,” Radiology **231**, 208–214 (2004).

[6] K. Moberg, N. Bjurstam, B. Wilczek, L. Rostgard, E. Egge, and C. Muren, “Computed assisted detection of interval breast cancers,” Eur. J. Radiol. **39**, 104–110 (2001).

[7] P. M. Taylor, J. Champness, R. M. Given-Wilson, H. W. W. Potts, and K. Johnston, “An evaluation of the impact of computer-based prompts on screen readers’ interpretation of mammograms,” Br. J. Radiol. **77**, 21–27 (2004).

[8] K. Hukkinen, T. Vehmas, M. Pamelo, and L. Kivisaari, “Effect of computer-aided detection on mammographic performance: Experimental study on readers with different levels of experience,” Acta Radiol. **47**, 257–263 (2006).

[9] T. W. Freer and M. J. Ulissey, “Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center,” Radiology **220**, 781–786 (2001).

[10] D. Gur, H. Sumkin, H. E. Rockette, M. Ganott, C. Hakim, L. A. Hardesty, T. S. W. R. Poller, R. Shah, and L. Wallace, “Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system,” J. Natl. Cancer Inst. **96**(3), 185–190 (2004).

[11] P. Taylor and R. M. Given-Wilson, “Evaluation of computer-aided detection (CAD) devices,” Br. J. Radiol. **78**, 26–30 (2005).

[12] R. L. Birdwell, P. Bandodkar, and D. M. Ikeda, "Computer-aided detection with screening mammography in a university hospital setting," Radiology **236**, 451–457 (2005).

[13] M. J. Morton, D. H. Whaley, K. R. Brandt, and K. K. Amrami, "Screening mammograms: Interpretation with computer-aided detection—Prospective evaluation," Radiology **239**(2), 375–383 (2006).

[14] R. M. Nishikawa and M. Kallergi, "Computer-aided detection, in its present form, is not an effective aid for screening mammography," Med. Phys. **33**(4), 811–814 (2006).

[15] E. A. Krupinski, "Computer-aided detection in clinical environment: Benefits and challenges for radiologists," Radiology **231**, 7–9 (2004).

[16] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1349–1380 (2000).

[17] H. Müller, A. Rosset, A. Garcia, J.-P. Vallée, and A. Geissbuhler, "Benefits of content-based visual data access in radiology," Radiographics **25**, 849–858 (2005).

[18] M. W. Vannier and R. M. Summers, "Sharing Images," Radiology **228**, 23–25 (2003).

[19] G. D. Tourassi, R. Vargas-Voracek, and C. E. Floyd, Jr., "Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information," Med. Phys. **30**, 2123–2139 (2003).

[20] G. D. Tourassi and C. E. Floyd, Jr., "Computer-assisted diagnosis of mammographic masses using an information-theoretic image retrieval scheme with BIRADs-based relevance feedback," Proc. SPIE **5370**, 810–816 (2004).

[21] H. Alto, R. M. Rangayyan, and J. E. L. Desautels, "Content-based retrieval and analysis of mammographic masses," J. Electron. Imaging **14**(2), 023016 (2005).

[22] I. El-Naqa, Y. Y. Yang, N. P. Galatsanos, R. M. Galatsanos, and M. N. Wernick, "A similarity learning approach to content-based image retrieval: Application to digital mammography," IEEE Trans. Med. Imaging **23**(10), 1233–1244 (2004).

[23] M. O. Honda, P. M. A. Marques, and J. A. H. Rodrigues, "Content-based image retrieval in mammography: Using texture features for correlation with BI-RADS categories," *Proceedings of the 6th International Workshop on Digital Mammography*, Bremen, Germany, June 22–25, 401–403 (2002).

[24] B. Zheng, A. Lu, L. A. Hardesty, J. H. Sumkin, C. M. Hakim, M. A. Ganott, and D. Gur, "A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment," Med. Phys. **33**(1), 111–117 (2006).

[25] C. Muramatsu, O. Li, K. Suzuki, R. A. Schmidt, J. Shiraishi, G. M. Newstead, and K. Doi, "Investigation of psychophysical measure for evaluation of similar images for mammographic masses: Preliminary results," Med. Phys. **32**(7), 2295–2304 (2005).

[26] P. M. Azevedo-Marques and M. O. Honda, "Content-based image retrieval in mammography: Using texture features for correlation with BI-RADS categories," Radiology **221**(Suppl. S), 161–162 (2001).

[27] C.-H. Wei, C.-T. Li, and R. Wilson, "A general framework for content-based medical image retrieval with its application to mammograms," Proc. SPIE **5748**, 134–143 (2005).

[28] Y. Rubner, J. Puzicha, J. M. Bhumann, and C. Tomasi, "Empirical evaluation of dissimilarity measures for color and texture," Comput. Vis. Image Underst. **84**, 25–43 (2001).

[29] S. Santani and R. Jain, "Similarity measures," IEEE Trans. Pattern Anal. Mach. Intell. **29**(9), 871–883 (1999).

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[31] J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes, *Medical Image Registration* (CRC, Boca Raton, FL, 2000).

[32] W. M. Wells, P. V. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," Med. Image Anal **1**, 35–51 (1996).

[33] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multi-modal image registration by maximization of mutual information," IEEE Trans. Med. Imaging **16**, 187–198 (1997).

[34] W. Li, "Mutual information functions versus correlation functions," J. Stat. Phys. **60**, 823–837 (1990).

[35] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based feature distributions," Pattern Recogn. **29**(1), 51–59 (1996).

[36] J. Puzicha, T. Hofmann, and J. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerta Rico, June 17–19, 267–272 (1997).

[37] M. Heath *et al.*, "Current status of the digital database for screening mammography," in *Digital Mammography* (Kluwer, Dordrecht, 1998). Available: http://marathon.csee.usf.edu/Mammography/Database.html.

[38] D. M. Catarious, A. H. Baydush, and C. E. Floyd, Jr., "A mammographic mass CAD system incorporating features from shape, fractal, and channelized Hotelling observer measurements: Preliminary results," Proc. SPIE **5032**, 111–119 (2003).

[39] D. M. Catarious, A. H. Baydush, and C. E. Floyd, Jr., "Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system," Med. Phys. **31**(6), 1512–1520 (2004).

[40] J. C. Russ, *The Image Processing Handbook* (CRC, Boca Raton, FL, 1995).

[41] D. Marr, *Vision* (Freeman, San Francisco, 1982).

[42] D. Catarious, A. Baydush, and C. E. Floyd, Jr., "The influence of true positive detection definitions on the performance of a mammographic mass CAD system," Med. Phys. **30**(6), 1368–1368 (2003).

[43] N. A. Obuchowski, "Receiver operating characteristic curves and their use in radiology," Radiology **229**, 3–8 (2003).

[44] D. B. Russakof, C. Tomasi, T. Rohlfing, and C. R. Maurer, Jr., "Image similarity using mutual information of regions," *The Eighth European Conference on Computer Vision, ECCV*, Prague, Czech Republic, May 11–14, 596–607 (2004).

[45] D. Rueckert, M. J. Clarkson, D. L. G. Hill, and D. J. Hawkes, "Non-rigid registration using higher-order mutual information," Proc. SPIE **3979**, 438–447 (2000).

[46] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Image registration by maximization of combined mutual information and gradient information," IEEE Trans. Med. Imaging **19**, 809–814 (2000).

Information-theoretic CAD system in mammography: Entropy-based indexing for computational efficiency and robust performance

Georgia D. Tourassi[a)] and Brian Harrawood

Digital Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, North Carolina 27705

Swatee Singh and Joseph Y. Lo

Digital Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, North Carolina 27705 and Department of Biomedical Engineering, Duke University, Durham, North Carolina 27710

(Received 25 January 2007; revised 1 May 2007; accepted for publication 30 May 2007; published 13 July 2007)

We have previously presented a knowledge-based computer-assisted detection (KB-CADe) system for the detection of mammographic masses. The system is designed to compare a query mammographic region with mammographic templates of known ground truth. The templates are stored in an adaptive knowledge database. Image similarity is assessed with information theoretic measures (e.g., mutual information) derived directly from the image histograms. A previous study suggested that the diagnostic performance of the system steadily improves as the knowledge database is initially enriched with more templates. However, as the database increases in size, an exhaustive comparison of the query case with each stored template becomes computationally burdensome. Furthermore, blind storing of new templates may result in redundancies that do not necessarily improve diagnostic performance. To address these concerns we investigated an entropy-based indexing scheme for improving the speed of analysis and for satisfying database storage restrictions without compromising the overall diagnostic performance of our KB-CADe system. The indexing scheme was evaluated on two different datasets as (i) a search mechanism to sort through the knowledge database, and (ii) a selection mechanism to build a smaller, concise knowledge database that is easier to maintain but still effective. There were two important findings in the study. First, entropy-based indexing is an effective strategy to identify fast a subset of templates that are most relevant to a given query. Only this subset could be analyzed in more detail using mutual information for optimized decision making regarding the query. Second, a selective entropy-based deposit strategy may be preferable where only high entropy cases are maintained in the knowledge database. Overall, the proposed entropy-based indexing scheme was shown to reduce the computational cost of our KB-CADe system by 55% to 80% while maintaining the system's diagnostic performance. © *2007 American Association of Physicists in Medicine*. [DOI: 10.1118/1.2751075]

Key words: computer-aided detection, mammography, image retrieval, mutual information entropy indexing

I. INTRODUCTION

The development and clinical application of computer-assisted detection (CADe) technology in mammography is a mature field of research with numerous published studies. Sampat *et al.* recently presented a review on the topic.[1] In addition, several commercial CADe products are available and in daily use. Nevertheless, the current performance of CADe is less than desired, especially with respect to the detection rate of breast masses as well as the CADe specificity. For example, the reported mass sensitivity of commercial CADe systems varies between 65% and 90% with 0.5 false positive detections per image.[2–4] The false positive rate is considered the main reason clinicians often distrust CADe aids. In the early development phase of this technology, it was assumed that the radiologists would be able to differentiate easily between visual cues that correspond to true abnormalities rather than false alarms. However, later studies showed that this assumption is not necessarily true.[5,6] Recognizing CADe false alarms proved to be a far more challenging task, heavily dependent on each radiologist's experience and attitude toward computer aids. Therefore, it is not surprising that several studies measuring the true clinical impact of CADe reported contradictory findings.[7–15] Researchers continue to improve upon CADe technology by addressing its current limitations. The most recent trends of CADe research include techniques that rely on the fusion of diverse detection schemes,[16] techniques that capitalize on the combination of mammographic views,[17–20] as well as techniques that address the human-computer communication aspects.[21–23]

The majority of existing CADe technology employs rule-based systems and artificial neural networks to make the final decision regarding the presence or absence of an abnormality. Recently knowledge-based (KB) systems were introduced as a possible alternative.[24,25] KB systems are designed to provide evidence-based decision support by comparing an unknown query case with known cases stored in a knowledge database. The main advantage of KB-CADe systems is their ability to capitalize on accumulating image data without requiring painstaking retraining or recalibration. Therefore, KB-CADe systems are inherently adaptive and new clinical cases can be continuously deposited in the knowledge database without interfering with the system's operation. Unfortunately, the clinical utilization of KB-CADe systems in mammography can be rather challenging due to the computational demands of maintaining and querying a continuously growing databank of mammograms.

In the past, we presented our own KB-CADe system for the detection[25] and diagnosis[26] of masses in screening mammograms, as well as its clinical application for further reduction of false positives generated by another CADe scheme.[27] Our system utilizes information-theoretic similarity measures to assess the relevance of a query case with those stored in its knowledge database. These similarity measures are based on the concept of image entropy as defined in information theory.[28] More importantly, the information-theoretic similarity measures are computed directly from the images without image preprocessing, mass segmentation, or feature extraction analysis. Although we have explored various entropy-based similarity measures, our latest study confirmed that by far the most effective measure is the mutual information and its normalized version.[27] Mutual information (MI) is a statistical measure of the amount of information redundancy between two images.[28] However, mutual information is computationally demanding. Blindly comparing the query case with every case stored in the knowledge database is a computationally expensive proposition, and it becomes impractical as more cases are deposited in the knowledge database.

There are two major aims in the study. They are both designed to determine whether we can incorporate an indexing strategy to improve the efficiency of the KB-CADe system without compromising its diagnostic performance. Our first aim is to apply the indexing strategy for searching effectively the knowledge database. Ideally, instead of comparing the query to all mammographic cases stored in the knowledge database, we would like to quickly identify a subset of cases that are potentially the most relevant cases to the specific query. A detailed analysis that involves mutual information could be restricted only to this subset. Our second aim is to use the indexing strategy as a selection mechanism to discard superfluous cases and build a concise knowledge database that contains only the most globally informative mammographic cases that are useful for a wide range of queries. Achieving both aims will lead to an intelligent KB-CADe system that balances diagnostic performance, computational speed, and database storage efficiency.

The paper is organized as follows. In subsection II A we provide an overview of our KB-CADe system; its fundamental components and general philosophy. Subsection II B describes the proposed entropy-based indexing modification to the system to facilitate faster analysis without compromising the system's overall diagnostic performance. Subsection II C describes the datasets involved in the study while subsection II D outlines the overall study design. Results are presented in Section III depending on the underlying aim. A summarization of the study findings, the possible implications for development of clinical knowledge-based systems in radiology, and future extensions of this work are discussed in the final section.

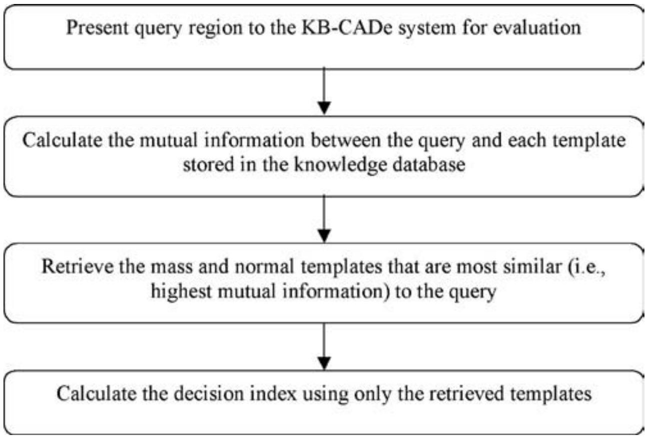

Present query region to the KB-CADe system for evaluation
Calculate the mutual information between the query and each template stored in the knowledge database
Retrieve the mass and normal templates that are most similar (i.e., highest mutual information) to the query
Calculate the decision index using only the retrieved templates

FIG. 1. Schematic of the KB-CADe system. The analysis is performed on a 512×512 pixel region extracted around a mammographic location that is recommended for evaluation by either a radiologist or an automated detection algorithm.

II. MATERIALS AND METHODS

A. The KB-CADe system: An overview

The major steps of our KB-CADe system have been described in detail before.[25,27] They are also shown as a block diagram in Fig. 1. The system is designed to provide an evidence-based second opinion for any mammographic location indicated by either a radiologist or an automated prescreening algorithm. A fixed size region of interest (ROI) is extracted around the indicated mammographic location. Then, the ROI is compared with a large knowledge database of mammographic templates. These templates are essentially mass and normal ROIs extracted from mammograms with known ground truth. The comparison between the query ROI Q and each template T stored in the knowledge database is based on the mutual information (MI). The MI is calculated as follows:

$$\mathrm{MI}(Q,T)=\sum_{q}\sum_{t}P_{QT}(q,t)\log_2\frac{P_{QT}(q,t)}{P_Q(q)P_T(t)}, \tag{1}$$

where $P_{QT}(q,t)$ is the joint probability density function of the two images based on their corresponding pixel values. $P_Q(q)$ and $P_T(t)$ are the marginal probability density functions. The above probability density functions are estimated from the image histograms.[29]

The stored templates are rank ordered according to their mutual information with the query case. Templates with higher MI are considered more similar to the query than templates with lower MI. The final decision regarding the query case is based on how similar the query is to the mass templates versus the normal templates retrieved from the knowledge database. Specifically, the decision index $D(Q)$ is expressed as

$$D(Q) = \frac{1}{k}\sum_{j=1}^{k} \mathrm{MI}(Q,M_j) - \frac{1}{k}\sum_{j=1}^{k} \mathrm{MI}(Q,N_j), \tag{2}$$

where M_j and N_j $(1 \leqslant j \leqslant k)$ are mass and normal templates that are retrieved from the knowledge database as the most similar to the query. Theoretically, as the decision index increases, the probability that the query case depicts a true mass should increase as well.

Our previous studies[25,27] showed that as the number k of retrieved templates increases, so does the diagnostic accuracy of the system. After a certain number has been retrieved, the diagnostic performance of the system plateaus. Thus, including more templates in the calculation of the decision index has no beneficial or detrimental effect. The implication of this finding is that essentially the KB-CADe system is equally effective by skipping the rank-ordering step and using all stored templates for the calculation of the decision index.

Note, however, that regardless of the minimum number of retrieved templates required for optimized performance, the computational complexity of the KB-CADe system remains essentially the same. The similarity is measured based on mutual information. Therefore, given a query ROI, the mutual information between the query and each stored template needs to be calculated first. Then, the stored templates can be rank ordered so that the k closest mass and the k closest normal templates are identified to derive the decision index. Although only those $2k$ templates are required for optimized decision making, the system still needs to perform as many MI calculations as the number of templates stored in the knowledge database. This is precisely the reason why including all available knowledge templates in the calculation of the decision index is a desirable alternative. The system's computational complexity remains the same, yet there is no need for careful optimization of the parameter k.

B. Entropy-based indexing

The construction of the KB-CADe decision [shown in Eq. (2)] requires that, given a query case Q, all $\mathrm{MI}(Q,T_i)$ calculations are executed and the stored templates T_i are rank ordered according to their similarity with the query. Theoretically, the effectiveness of KB decision systems depends on the comprehensiveness of the knowledge database. As new and more diverse templates are stored in the database, KB decisions tend to become more accurate. We have observed the same in a preliminary study performed with our own KB-CADe system.[30] In contrast, as the knowledge database increases in size, the computational efficiency of the system declines since the search and rank-ordering steps take a substantially longer time. Therefore, the exhaustive search for similar templates becomes impractical, compromising the system's ability to provide real time second opinions. This is a serious limitation, particularly for our system, because mutual information is a computationally demanding similarity measure.

We propose an entropy-based indexing scheme to address this limitation. The indexing scheme operates as follows. Instead of comparing the query case with the whole knowledge database, we restrict the comparisons to those stored templates that share the same level of image complexity as the query case. The image complexity is measured using image entropy (H), a measure of randomness of the gray level distribution in the image:

$$H = -\sum_{x} p(x)\log_2(p(x)), \tag{3}$$

where the probability $p(x)$ that an image pixel will have the intensity value x is estimated from the image histogram. Although other indices could be employed for the same purpose, entropy is a very attractive and logical choice for our own KB-CADe system. Not only the template entropies can be easily calculated, stored, and rank ordered in the knowledge database, but also measuring the query's entropy is already a necessary step for the calculation of the MI similarity measure. We will elaborate on this point.

According to information theory, MI measures how much the uncertainty (or entropy) of the query case is reduced if we have prior knowledge about the template.[28] Thus, MI is the conditional entropy of the query case given the template:

$$\mathrm{MI}(Q,T) = H(Q|T) = H(Q) + H(T) - H(Q,T). \tag{4}$$

If the query and the template are independent, then knowing the template does not really help draw any conclusions about the query. Their mutual information is 0, and therefore the uncertainty about the query remains unchanged. If, on the other hand, the query is identical to the template, then knowing the template provides complete knowledge about the query. In this case, their mutual information is maximized, and the uncertainty of the query given prior knowledge of the template is reduced to 0.

Integrating the entropy-based indexing scheme in the existing KB-CADe system does not require any major additional calculations. Essentially, the query entropy H(Q) is already part of the MI calculations that are necessary for decision making [as shown in Eq. (4)]. Furthermore, the template entropy H(T) can be calculated offline and stored as soon as each template is deposited in the database. Figure 2 illustrates the block diagram of the KB-CADe system enhanced with the entropy-based indexing scheme.

For this study, a nearest-neighbor clustering implementation of the entropy-based indexing scheme was applied to identify the intermediate set of "most relevant" templates. Given a query case, only a fixed number K of mass and normal templates are retrieved for further analysis. The retrieved templates are the ones that are closest to the query in terms of their individual entropies. A separate search is performed among the mass and normal templates to retrieve an equal number $(K/2)$ of mass and normal templates for a total of K retrievals. The nearest-neighbor implementation is at-

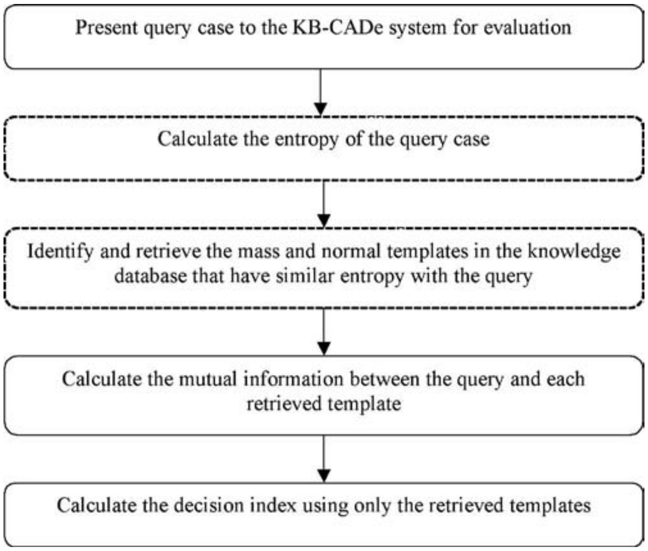

Present query case to the KB-CADe system for evaluation
Calculate the entropy of the query case
Identify and retrieve the mass and normal templates in the knowledge database that have similar entropy with the query
Calculate the mutual information between the query and each retrieved template
Calculate the decision index using only the retrieved templates

FIG. 2. Schematic of the KB-CADe system that incorporates the entropy-based indexing scheme for fast search of the knowledge database. The scheme assumes that every time a new template is stored in the knowledge database its entropy is calculated, and the template is indexed accordingly. The two additional steps that implement the intermediate, entropy-based retrieval of stored templates are shown in dash outline.

tractive because the computational cost per query is fixed for the specific computer configuration that runs the KB-CADe system.

C. Datasets

The study was performed using two datasets of different difficulty level. The datasets have been described in a previous paper.[27] We used the same datasets to facilitate a detailed comparison of the results and show the progression of this research. Both datasets included 512×512 pixel ROIs extracted from mammographic cases from the Digital Database of Screening Mammography (DDSM).[31] The mammograms were selected from the cancer, benign, and normal DDSM volumes digitized using the Lumisys scanner at 50 μm per pixel. The abnormal DDSM volumes included cases that required biopsy or additional diagnostic studies to establish the ground truth. Therefore, our datasets did not include any "benign-without-callback" cases since radiologists would not really request a second opinion on such less challenging group of cases.

The first dataset contained 1,820 ROIs in total; 489 with malignant mass, 412 with benign mass, and 919 normal. The normal ROIs were selected from both normal and abnormal mammographic cases as long as the imaged breast did not contain any physician annotations in either mammographic view. The mass ROIs were extracted around the DDSM physician's annotation. If the annotated mass was close to the breast skin line, the extracted ROI extended beyond the breast skin line, thus covering air pixels. There were 42 such mass ROIs in the first dataset. The normal ROIs were selected randomly within the breast region.

The second dataset contained 483 ROIs extracted from 100 DDSM/Lumisys cases completely different from those used to create the first dataset. Of those, 84 ROIs depicted true masses (44 malignant and 40 benign) and 399 ROIs were false positive detections. There were four true masses located close to the breast skin line. The false positive ROIs were extracted around normal mammographic locations that were indicated as suspicious by a feature-based CADe system developed before in our laboratory.[32,33] Since the normal ROIs in Dataset 2 have masslike characteristics, it is expected that the discrimination of mass from normal ROIs is substantially harder in Dataset 2 than Dataset 1.

To facilitate a better assessment of the difficulty level of the mass detection task for both datasets, all masses were furthered indexed according to their subtlety rating provided in the DDSM. The mass subtlety rating is a subjective rating provided by the DDSM radiologist regarding lesion visibility. This rating ranges from 1 to 5 where a rating of 5 indicates that a lesion is 5 times more obvious than a lesion with rating of 1. Specifically, in Dataset 1 there were 23 (2.6%) masses with rating 1, 64 (7.1%) with rating 2, 151 (16.8%) with rating 3, 204 (22.6%) with rating 4, and 459 (50.9%) with rating 5. A similar subtlety distribution was observed for the masses present in Dataset 2. There were 1 mass (2.3%) with rating 1, 6 (7.1%) with rating 2, 18 (21.4%) with rating 3, 16 (19%) with rating 4, and 43 (51.2%) with rating 5.

D. Evaluation studies

The two study aims were pursued separately with two different experiments. The first experiment was designed to answer the question: "Can entropy-based indexing improve the speed of search and decision making for any given query?" Thus, the first experiment evaluated the modified KB-CADe system shown in Fig. 2. Different values of the nearest neighbor parameter K were investigated to determine its optimal value.

The second experiment was designed to answer: "Can entropy-based indexing reduce the size of the knowledge database by discarding less useful templates without compromising the overall diagnostic performance of the system?" This experiment does not target computational efficiency per se but rather system efficiency with respect to data storage and maintenance. For the second experiment, the available ROIs were first ranked according to their entropy. The ranking was performed separately for each class of templates (i.e., mass and normal). Then, the performance of the KB-CADe system was monitored starting with an equal number of the highest-entropy mass and normal templates in its knowledge database and adding progressively lower-entropy templates from each class. We call this database reduction scheme "high-entropy" selection strategy. The same experiment was repeated by depositing the lowest-entropy mass and normal templates in the database first and then adding progressively an equal number of higher-entropy templates from each class. We call this database reduction scheme "low entropy."

Both experiments were initially performed using Dataset 1 and a leave-one-case-out sampling scheme. Specifically,

TABLE I. Entropy-based statistics for the two study datasets

Type of ROIs	Number of ROIs	Entropy ($\mu \pm \sigma$)
Dataset 1: Malignant mass	489	5.46±0.56
Dataset 1: Benign mass	412	5.49±0.56
Dataset 1: Normal	919	5.72±0.24
Dataset 2: Malignant mass	44	5.36±0.69
Dataset 2: Benign mass	40	5.50±0.42
Dataset 2: Normal	399	5.41±0.77

each ROI in the dataset was excluded once as the query. Of the 1,819 ROIs remaining in Dataset 1, only those extracted from DDSM cases different than the query's served as the knowledge database. Excluding all ROIs coming from the same case as the query eliminates any possible biases. The same experiments were also performed using Dataset 1 as the knowledge database and Dataset 2 as the test bed for additional validation on a clinically more challenging task; the discrimination of true masses from false positive findings.

The results of all experiments were analyzed using receiver operating characteristic (ROC) analysis[34] with the KB-CADe decision index being the decision variable. The ROC analysis was performed with the ROCKIT software developed by Metz at the University of Chicago. Detection performance was measured using the overall (A_z)[34] and partial ROC (${}_{0.90}A_z$) area index.[35] The partial ROC area measures the average specificity of the KB-CADe system when it operates with sensitivity in the range of 90% to 100%. In cancer detection, the partial area index is clinically more relevant since the consequences of missing cancer are more severe than those of a false alarm.

III. RESULTS

A. Entropy-based data statistics

Initially, the entropy of all ROIs in our datasets was measured. Consistent with our previous studies, the entropy of each 512×512 pixel ROI was estimated using the histogram approach with 64 bins. Table I summarizes the statistics of the entropy index for both datasets and for each subgroup of cases (malignant mass, benign mass, normal).

Overall, the average entropy of the mass ROIs was statistically significantly lower than that of the normal ROIs for both datasets (p-value of a two-tailed t-test <0.001 at 95% confidence level). Despite this, utilizing entropy as the decision index to discriminate between mass and normal ROIs resulted in mediocre ROC performance ($A_z=0.70\pm0.01$). Thus, using entropy as the decision variable is not particularly useful for mass detection.

There was no statistically significant difference between the average entropy of the benign and malignant masses. This finding was true for both datasets (two-tailed p-value $=0.42$ for Dataset 1 and two-tailed p-value$=0.21$ for Dataset 2 at 95% confidence level) with corresponding ROC area indices of $A_z=0.53\pm0.02$ for Dataset 1, and $A_z=0.54\pm0.06$ for Dataset 2, respectively. Thus, entropy alone is not useful for diagnostic purposes either. Finally, no statistical difference was observed in the average entropy of masses between the two datasets (two-tailed p-value$=0.45$ at 95% confidence level). In contrast, the average entropy of the normal ROIs in Dataset 1 was statistically significantly higher than that of the normal ROIs in Dataset 2 (two-tailed p-value<0.001 at 95% confidence level). This finding is not really surprising. The normal ROIs included in Dataset 2 are suspicious, "mass-like" ROIs and not randomly chosen as in Dataset 1. Therefore, it is expected that their average entropy should be closer to that of the mass ROIs, as confirmed in Table I.

B. AIM 1: Entropy-based indexing for effective search of the knowledge database

The diagnostic performance of the KB-CADe system was evaluated without and with the entropy-based indexing scheme (Fig. 1 versus Fig. 2 configurations). The K-nearest neighbor implementation was investigated for K ranging from 2 (i.e., retrieve the one mass and the one normal template that are closest in entropy to the query) to the maximum possible value (i.e., include all available mass and normal templates). The results of this experiment are summarized in Fig. 3 for both datasets. The figure shows how the diagnostic performance of the system changes as the number K of nearest neighbor-in-entropy templates retrieved for further analysis increases. The diagnostic performance is reported in terms of the overall and partial ROC area indices [Figs. 3(a) and 3(b), respectively].

Since the ultimate goal of the entropy-based indexing scheme is to help KB-CADe achieve its best performance with the minimum possible number of computations per query, it is important to establish the baseline performance of the system. The baseline performance is the best performance observed when the system operates in its original configuration without any entropy-based indexing (shown in Fig. 1). For Dataset 1, the baseline performance was $A_z=0.87\pm0.01$, and ${}_{0.90}A_z=0.48\pm0.02$ for all masses. The system's performance was significantly higher for malignant than for benign masses with respect to both the overall ROC area index ($A_z=0.89\pm0.01$ versus $A_z=0.84\pm0.01$, two-tailed p-value<0.004), as well as the partial ROC area index (${}_{0.90}A_z=0.57\pm0.03$ versus ${}_{0.90}A_z=0.41\pm0.03$, two-tailed p-value<0.002). Using Dataset 1 as the knowledge database and Dataset 2 as the queries, the baseline ROC performance of the KB-CADe system was $A_z=0.81\pm0.03$ and ${}_{0.90}A_z=0.32\pm0.06$ for all masses. The detection performance was very similar for both benign ($A_z=0.81\pm0.04$, ${}_{0.90}A_z=0.32\pm0.09$) and malignant masses ($A_z=0.82\pm0.03$, ${}_{0.90}A_z=0.33\pm0.09$). Although there was a substantial performance decline from Dataset 1 to Dataset 2, this was an expected finding considering that the detection task is far more challenging in Dataset 2.

The computational demands of the KB-CADe system in its original configuration were proportional to the size of its knowledge database. For Dataset 1, approximately 1815–

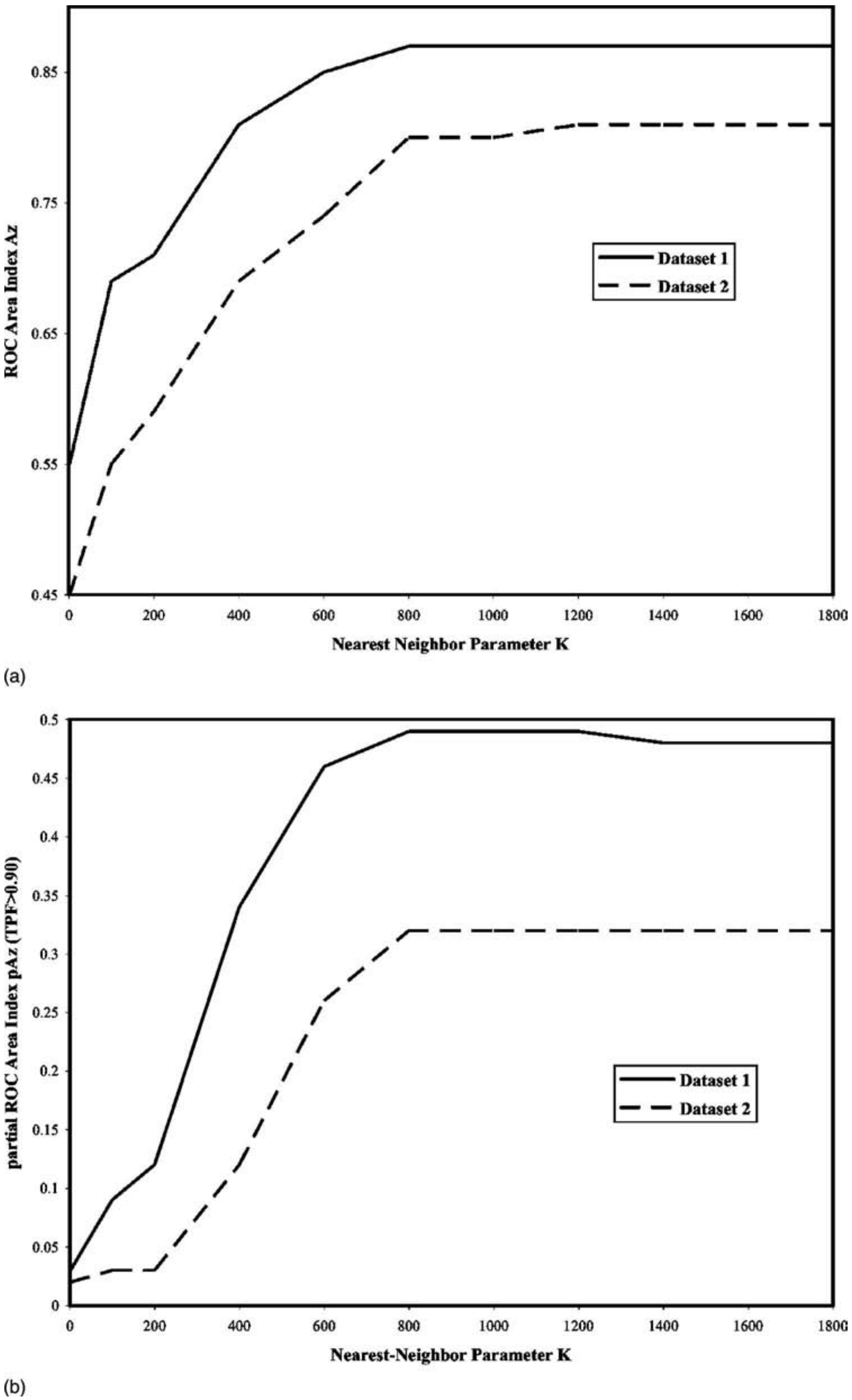


FIG. 3. Diagnostic performance of the KB-CADe system including the entropy-based indexing scheme based on the (a) overall and (b) partial ROC area indices. Performance is shown at steadily increasing values of the *K* nearest-neighbor parameter for Datasets 1 and 2.

1819 MI calculations were needed per query. This number is not fixed in Dataset 1 due to the leave-one-case-out sampling scheme utilized for this experiment. Some mammographic cases contributed more than one ROI in Dataset 1. When an ROI was excluded to serve as the query, all other ROIs extracted from the same mammographic case were excluded for similarity assessment to avoid a positive bias. Consequently, for those queries the knowledge database contained slightly fewer than 1819 (=1820-1) templates. However, there were no cases contributing more than five ROIs in Dataset 1. In contrast, 1820 MI calculations were performed when the system was tested on cases drawn from Dataset 2.

This number is fixed since there was no overlap between the knowledge database (Dataset 1) and the query database (Dataset 2).

Figure 3 shows that including the entropy-based indexing scheme did not result in any further improvement of the baseline KB-CADe diagnostic performance for Dataset 1. It was able however to duplicate the baseline performance. By limiting its MI-based calculations to only the 400 mass and 400 normal templates that are closest to the query (in entropy space), the KB-CADe maintained its baseline diagnostic performance in Dataset 1 (A_z=0.87±0.01 and ${}_{0.90}A_z$=0.49±0.03). A consistent trend was observed with Dataset 2. Using the entropy-based indexing scheme, the KB-CADe system performed robustly (A_z=0.80±0.03 and ${}_{0.90}A_z$=0.32±0.06) while reducing the number of computations from 1,820 to 800 per query. Overall, the nearest-neighbor entropy-based indexing scheme improved the computational efficiency of the system by almost 55%, reducing the number of necessary MI-based computations from roughly 1800 down to 800 calculations per query.

C. AIM 2: Entropy-based indexing for identifying informative templates

The entropy index was also evaluated as the foundation of a selection strategy to determine which templates could be eliminated from the knowledge database without compromising the overall diagnostic performance of the KB-CADe system. Trimming down the knowledge database so that only the globally most informative templates are preserved helps satisfy possible database storage limitations.

ROC performance is presented separately for Dataset 1 (based on the leave-one-case-out scheme) and then validated on Dataset 2 as done with aim 1. Figure 4(a) shows how the overall ROC area index of the KB-CADe system changes based on the composition of its knowledge database for Dataset 1. As a reference point, the figure also includes the performance of the system when templates are added in the knowledge database by selecting them randomly from the available pool of mass and normal templates. This database-building scheme is labeled “random selection.” Figure 4(b) summarizes the results of the same experiment using Dataset 2 as the testbed. The results shown for the random selection database-building scheme are based on averaging the KB-CADe performance across five different random samplings.

Some interesting trends are observed in Fig. 4(a). First, the ROC performance of the system is dramatically low when the knowledge database is sparse. As more templates are deposited in the database, the KB-CADe detection performance improves steadily. The most rapid improvement occurs with the high-entropy selection scheme. Depositing first the higher entropy mass and normal templates improves rapidly the diagnostic performance of the KB-CADe system. The system achieves its highest performance in Dataset 1 (A_z=0.89±0.01, ${}_{0.90}A_z$=0.48±0.03) using only the 300 mass and 300 normal templates with the highest entropy. Actually, this ROC performance is significantly higher than that achieved with the 400 highest entropy templates (A_z=0.87±0.01) or with more than 800 templates (A_z=0.88±0.01). However, no significant difference is observed with respect to the partial ROC area index. Overall, Fig. 4(a) suggests that maintaining a database of only 600 high-entropy templates instead of 1800 does not affect the system’s diagnostic performance. In contrast, the ROC performance of the system remains significantly lower when relying on low-entropy templates. As more higher-entropy templates are added in the knowledge database, the diagnostic performance of the system steadily improves. Surprisingly, random storing of templates works quite well. However, the random selection scheme never outperforms the results obtained when the KB-CADe system relies only on the higher-entropy templates. Furthermore at 600 templates, the high-entropy building scheme results in significantly better performance than the random selection plan with respect to the ROC area index (two-tailed p-value=0.02) but not with respect to the partial ROC area index (two-tailed p-value=0.06).

A similar trend was observed with Dataset 2 [Figure 4(b)]. Relying on the 600 templates with the highest entropy, the KB-CADe system performed very similar to the baseline performance (A_z=0.80±0.03, ${}_{0.90}A_z$=0.34±0.06). Using the 1000 higher entropy templates, the previous performance improved significantly with respect to the ROC area index (A_z=0.82±0.03, two-tailed p-value=0.006) but not with respect to the partial ROC area index (${}_{0.90}A_z$=0.33±0.06, two-tailed p-value=0.76). It should be noted, however, that this optimized performance did not reach statistical significance compared to random selection as we observed in Dataset 1. Actually, the random selection strategy appears to be a quite effective deposit strategy for the beginning stages of the database building process. This finding may be due to the difficulty of the detection task and the substantially smaller size of Dataset 2 compared to Dataset 1.

An examination of the 600 high-entropy templates revealed a similar distribution of mass and normal templates according to mammographic density. Specifically, 17.3% high-entropy mass templates came from fatty breasts, 59.3% from fibroglandular breasts, 21.3% from heterogeneous breasts, and 2% from dense breasts. Similarly, 13.7% high-entropy normal templates came from fatty breasts, 47.3% from fibroglandular breasts, 33% from heterogeneous breasts, and 6% from dense breasts. Of the 300 high-entropy mass templates, exactly 50% depicted malignant masses, and the remaining depicted benign masses. Furthermore, 1.3% of the mass templates were assigned a DDSM subtlety rating 1, 7% a subtlety rating 2, 15% a subtlety rating 3, 25% a subtlety rating 4, and 51.7% a subtlety rating 5. The subtlety rating distribution of the 300 high-entropy masses is very similar to the one reported earlier regarding the full set of 901 masses present in Dataset 1.

D. Putting it all together

The previous experiments suggested that entropy-based indexing could be a useful modification for our KB-CADe

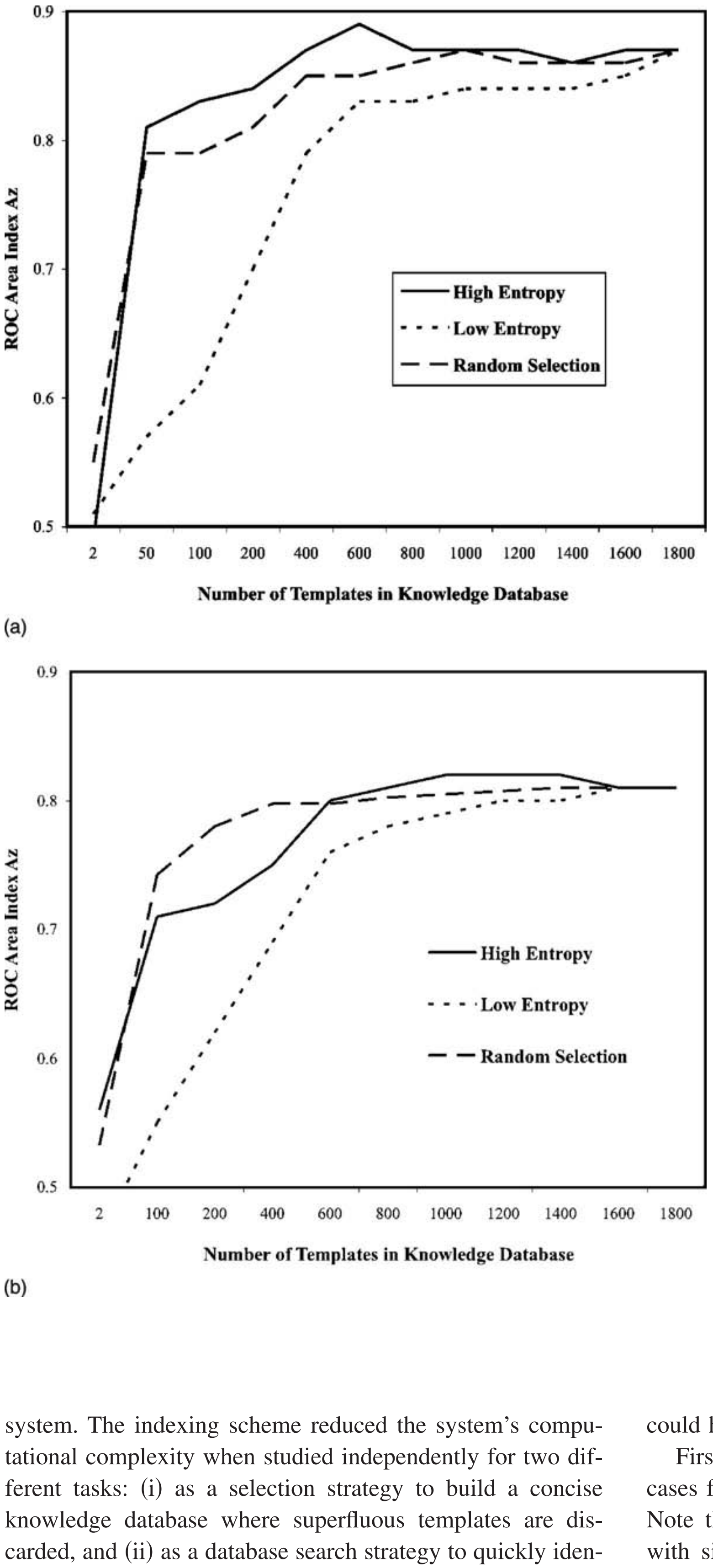


FIG. 4. ROC area index of the KB-CADe system based on the total number and type of templates stored in the knowledge databank. The reported results assume that the prevalence of mass and normal templates is consistently maintained at 50% in the knowledge database. Detection performance is shown with respect to the ROC area for (a) Dataset 1 (σ_{A_z}=0.01) and (b) Dataset 2 (σ_{A_z}=0.03).

system. The indexing scheme reduced the system's computational complexity when studied independently for two different tasks: (i) as a selection strategy to build a concise knowledge database where superfluous templates are discarded, and (ii) as a database search strategy to quickly identify a subset of stored templates to be used for decision making regarding the query. A logical extension of this work is to test whether putting both the entropy-based database reduction and entropy-based database search mechanisms in place could have any additional computational benefits.

First, the 600 most informative (i.e., highest entropy) cases from Dataset 1 were used as the knowledge database. Note that 300 were mass and 300 were normal templates with similar entropy statistics (average entropy 5.82±0.04 for mass versus 5.85±0.03 for normal cases). Then, the KB-CADe was tested on Dataset 2 for false positive reduction. During testing, entropy based indexing with the K nearest neighbor implementation was applied for decision making.

TABLE II. Application of the KB-CADe system enhanced with entropy-based database reduction and entropy-based database search mechanisms.

Number of nearest-neighbor templates K	Knowledge database composition (600 templates in total)			
	High-entropy templates		Randomly selected templates	
	A_z	${}_{0.90}A_z$	A_z	${}_{0.90}A_z$
100	0.79±0.03	0.29±0.06	0.70±0.03	0.20±0.05
200	0.79±0.03	0.27±0.06	0.76±0.03	0.22±0.06
400	0.81±0.03	0.32±0.06	0.77±0.03	0.25±0.06
600	0.81±0.03	0.33±0.06	0.79±0.03	0.27±0.06

The results of this study are summarized in Table II. The results are shown for representative values of the parameter K ranging from 100 to 600 (i.e., the full, high-entropy knowledge database).

Table II shows that KB-CADe achieves its highest performance ($A_z = 0.81 \pm 0.03$, ${}_{0.90}A_z = 0.32 \pm 0.06$) when focusing on the 400 out of the 600 high-entropy templates that are nearest in entropy to the query. Thus, for the specific experiment, entropy-based indexing results in almost 80% reduction in computations per query. Instead of calculating the MI between a query and the full database of 1820 available templates to make an optimized decision, only 400 MI calculations are necessary without any performance decline. Similar to the previous section, we include as a reference point the performance of the KB-CADe system when (i) its knowledge database is composed of 600 randomly selected templates (300 mass+300 normal) and (ii) the system is enhanced with entropy-based nearest-neighbor modification. The random selection scheme was repeated five times, selecting different templates from the available Dataset 1 to build a knowledge database of 600 templates. The results reported in the table for the random selection building scheme are based on averaging the KB-CADe performance across those five experiments.

The table clearly shows that combining the entropy-based search strategy with a smaller knowledge database of randomly selected templates is significantly less effective than when the entropy-based search is performed in a knowledge database of carefully tailored higher-entropy templates. The latter strategy produced statistically significantly better results with respect to both performance indices (two-tailed p-value < 0.05) for $K = 400$.

The best performance achieved for $K = 400$ with the entropy-enhanced KB-CADe system was further analyzed with respect to mass subtlety to assess whether the proposed entropy modification impacts system performance differently depending on mass visibility. Table III shows the detection performance of the KB-CADe system when it operates without and with the entropy modification. The table confirms that performance remains robust across the original and the modified system for all mass subtlety ratings. This finding is consistent with respect to both the overall and partial ROC area indices. A noticeable decline in performance was observed for masses with a subtlety rating of 3. However, the difference was not statistically significant (two-tailed p-value $= 0.1453$). Actually, none of the differences were statistically significant at the 95% confidence level. Therefore, the entropy-based modification maintains the detection performance of the original system irrespective of mass subtlety. Note that masses with subtlety ratings 1 and 2 were combined into one group since there was only one mass with subtlety rating 1 in Dataset 2.

IV. DISCUSSION

The computational complexity of any knowledge based system depends on two factors: (i) the size of its knowledge database and (ii) the numerical demands of the pairwise

TABLE III. Detection performance according to the mass subtlety rating for the original and the entropy-modified KB-CADe system.

Mass subtlety	ROC area index A_z		Partial ROC area index ${}_{0.90}A_z$	
	Original KB-CADe	Entropy-modified KB-CADe	Original KB-CADe	Entropy-modified KB-CADe
1+2 (subtle)	0.87±0.05	0.91±0.03	0.60±1.6	0.76±0.09
3	0.82±0.04	0.76±0.06	0.46±0.11	0.29±0.12
4	0.81±0.05	0.80±0.05	0.41±0.12	0.43±0.12
5 (obvious)	0.81±0.04	0.81±0.04	0.24±0.09	0.26±0.09

comparisons between the query and each case stored in the knowledge database. As KB systems become increasingly more popular for medical decision making in radiology, the practical limitations of maintaining and using these systems need to be considered carefully.

We have studied this issue with respect to our own KB-CADe system for the detection of masses in screening mammograms. Since our system utilizes mutual information, a featureless similarity measure that is computed directly from the image histograms, the system is spared the elaborate image preprocessing steps of feature-based KB-CADe systems. However, the computational complexity of our KB-CADe still increases much faster than that of a feature-based CADe system. An exhaustive calculation of the mutual information between a query case and every other case stored in the knowledge database becomes impractical as more new cases are deposited in the database. Furthermore, continuous deposit and accessibility of mammographic cases will become impractical in the long run, due to data storage requirements.

To address these limitations, we have proposed an entropy-based indexing scheme as an effective way to improve the efficiency of our KB-CADe system while not reducing its overall detection performance. The proposed entropy-based indexing scheme was evaluated first as a search mechanism to sort through the available data fast and identify the stored cases that are more diagnostically useful for a specific query. In this capacity, the entropy-based indexing scheme was applied to improve the speed of analysis and computation time per query. In addition, the same indexing scheme was evaluated as a selection mechanism to maintain the globally most useful cases in the knowledge database. Although this selection mechanism does not really affect the computational efficiency of our KB-CADe system, it does avoid excessive storage of cases that are superfluous. Case reduction is often necessary in large knowledge databases with a fixed storage limit to avoid needless storage of cases that do not improve diagnostic performance. Both case reduction and search mechanisms are critical components for efficient clinical knowledge-based systems. They facilitate easier maintenance and navigation of knowledge databases. It should be noted that although our present study relied only on a single attribute (i.e., image entropy), the proposed indexing scheme could be easily modified to include other image and/or textual descriptors of clinical importance.

First, the proposed indexing scheme was shown to improve the speed of search of the knowledge database by 55%, while not compromising the detection performance of the system. This result was confirmed with two datasets of different difficulty levels. Specifically for a typical query case, the KB-CADe system running on a single processor of an Apple Power Mac G5 (2×2.7 GHz POWERPC CPU with 8 GB memory) requires 70 s for 1820 comparisons without entropy-based indexing. Integrating the entropy-based indexing scheme reduces the computational time down to 31 s per query. The entropy-based sorting and searching step adds only 2 s in the decision making process (for a total of 31 s). It should be emphasized that significant gains are achieved with parallel processing. For example, running the KB-CAD system on ten processors reduces the computational demands by tenfold.

These results were based on a nearest-neighbor clustering implementation. Actually, we also explored two other implementation schemes that proved to be significantly less effective. The other two schemes were based on absolute entropy distance or case-dependent distance from the query case. The fixed-distance implementation assumes that there is a fixed threshold e that is optimal for all queries. Therefore, only templates with entropy within a fixed distance e from the query's entropy are retrieved for further analysis. The case-dependent distance implementation assumes that the distance is a fraction of the query's entropy (e.g., retrieve templates with entropy that is within 10% from that of the query's). Therefore, low-entropy (or low uncertainty) queries require a tighter radius for retrieving similar templates than the higher-entropy queries. None of these two implementations were able to achieve similar diagnostic performance while reducing significantly the computational burden of the KB-CADe system. We are currently working on an evolutionary programming technique to optimize the case-dependent distance implementation scheme. We believe that a case-dependent clustering scheme is a more promising strategy because it takes into account the unique characteristics of each query case.

Second, a dramatic improvement in data storage requirements was observed when the entropy-based indexing scheme was asked to potentially eliminate knowledge cases that do not seem to contribute much to the overall performance of the system. The study suggested that the higher-entropy knowledge cases are more useful for maintaining the expected performance level of our system. The system's performance was optimized when relying only on the 300 mass and 300 normal knowledge cases with the highest entropy. For the specific datasets employed in this study, this finding translates into 66% reduction of data storage requirements for our KB-CADe system.

The higher diagnostic contribution of the high-entropy mass and normal templates is not very surprising. In medical imaging, we often consider entropy a statistical measure of randomness that is used to characterize image texture. Thus, our study suggests that mass templates and normal templates with more complicated texture seem to be more useful for the overall detection performance of our KB-CADe system. A detailed analysis of our results with Dataset 1 showed that this finding was consistent for both high-entropy and low-entropy query cases. Specifically, for low-entropy queries a knowledge database composed of either low- or high-entropy templates was equally effective ($A_z=0.77\pm0.02$). Furthermore, this performance was very similar to that achieved relying on the full database of all available templates ($A_z=0.76\pm0.02$). In contrast, for high-entropy queries, a knowledge database composed of only high-entropy templates was significantly more effective than a database containing only low-entropy templates ($A_z=0.90\pm0.01$ versus $A_z=0.83\pm0.02$). As with the low-entropy queries, relying on the full knowledge database was as effective for high-entropy queries ($A_z = 0.89 \pm 0.02$). However, a noticeable difference in KB-CADe performance is observed between low- and high-entropy queries regardless of the knowledge database composition. Further investigation is needed to explain this finding, as well as identify specific queries for which relying only on high-entropy knowledge cases has a detrimental effect. Such data mining is essential to derive an effective case deposit mechanism, after the basic body of knowledge has been built in the database.

An unexpected finding of the study was that the random selection strategy was quite effective for building the knowledge database. This was particularly true in the beginning stages of the knowledge database building process. Our results suggest that while initially a random selection scheme is sufficient to build a certain body of knowledge in the database, a more sophisticated selection strategy maybe preferable later to determine whether an incoming template should be deposited or not. Extensive studies with diverse datasets are needed to validate the generalizability of the above observations.

By pursuing separately the two study aims we were able to delineate the contribution of the entropy-based indexing scheme as a database search and database reduction mechanism. The indexing scheme was as effective for building a concise knowledge database as it was for searching the database to find templates that are diagnostically useful for a specific query. Putting both mechanisms in place resulted in improvement regarding both computational speed and data storage requirements. Based on the results reported in Tables II and III, the computational demands were reduced by almost 80% per query case while the diagnostic performance of the KB-CADe was effectively maintained irrespective of mass subtlety. Future analysis will focus on delineating the impact of mass size as well, to elucidate the impact of the entropy-based system modification, system limitations, as well as possible extensions for the detection of the most challenging, smaller masses. Since image entropy is calculated using the full ROI and not just the suspected mass, the contribution of the background is substantial for smaller masses. Therefore, it is possible that an entropy-based indexing scheme using only the segmented abnormality is a better approach. Unfortunately, the DDSM database does not include information regarding the size of the annotated masses. The DDSM-provided annotations cannot be utilized for mass size estimation because they are often substantially larger than the actual masses. For the present study, we used the DDSM mass subtlety rating as a surrogate measure of case difficulty in place of mass size.

The major limitation of our study is that it is empirical in nature. Therefore, it is expected that the conclusions depend on the available data. This is precisely the reason we employed two different datasets. Although both datasets were generated using mammographic cases from the same benchmark database of screening mammograms, the selection criteria were different between the datasets to simulate two progressively more challenging detection tasks. It was reassuring to observe the same general trends between the two datasets.

In conclusion, it is important to balance diagnostic performance, computational speed, and data storage requirements when developing knowledge databases for CADe use. Our entropy-based indexing scheme was a significant step toward achieving those goals with our own knowledge-based CADe system in mammography.

ACKNOWLEDGMENTS

This work was supported by grants R01 CA101911 and R01 CA112437 from the National Cancer Institute and by grant W81XWH-05-1-0293 from the Army Breast Cancer Research Program.

[a] Author to whom correspondence should be addressed. Electronic mail: georgia.tourassi@duke.edu

[1] M. P. Sampat, M. K. Markey, and A. C. Bovik, “Computer-aided detection and diagnosis in mammography,” in *Handbook of Image and Video Processing*, 2nd ed. (Academic, New York, 2005), pp. 1195–1217.

[2] A. Malich, C. Marx, M. Facius, T. Boehm, M. Fleck, and W. A. Kaiser, “Tumour detection rate on a new commercially available computer aided detection system,” Eur. Radiol. **11**, 2454–2459 (2001).

[3] S. Ciatto, M. R. Del Turco, G. Risso, S. Catarzi, R. Bonardi, V. Viterbo, P. Gnutti, B. Giglielmoni, L. Pinelli, A. Pandiscia, F. Navarra, A. Lauria, R. Palmiero, and P. L. Indovina, “Comparison of standard reading and computer aided detection (CAD) on a national proficiency test of screening mammography,” Eur. J. Radiol. **45**, 135–138 (2003).

[4] D. Gur *et al.*, “Computer-aided detection performance in mammographic examinations of masses: Assessment,” Radiology **233**, 418–423 (2004).

[5] B. Zheng *et al.*, “Soft-copy mammographic readings with different computer-assisted detection cuing environments: Preliminary findings,” Radiology **221**, 633–640 (2001).

[6] B. Zheng *et al.*, “Detection and classification performance levels of mammographic masses under different computer-aided detection cueing environments,” Acad. Radiol. **11**, 398–406 (2004).

[7] L. J. W. Burhenne *et al.*, “Potential contribution of computer-aided detection to the sensitivity of screening mammography,” Radiology **215**, 554–562 (2000).

[8] R. L. Birdwell, D. M. Ikeda, K. F. O’Shaughnessy, and E. A. Sickles, “Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection,” Radiology **219**, 192–202 (2001).

[9] R. F. Brem *et al.*, “Improvement in sensitivity of screening mammography with computer-aided detection: A multi-institutional trial,” Am. J. Roentgenol. **181**, 687–693 (2003).

[10] M. A. Helvie *et al.*, “Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection: pilot clinical trial,” Radiology **231**, 208–214 (2004).

[11] K. Hukkinen, T. Vehmas, M. Pamelo, and L. Kivisaari, “Effect of computer-aided detection on mammographic performance: Experimental study on readers with different levels of experience,” Acta Radiol. **47**, 257–263 (2006).

[12] D. Gur, H. Sumkin, H. E. Rockette, M. Ganott, C. Hakim, L. A. Hardesty, W. R. Poller, R. Shah, and L. Wallace, “Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system,” J. Natl. Cancer Inst. **96**, 185–190 (2004).

[13] P. Taylor and R. M. Given-Wilson, “Evaluation of computer-aided detection (CAD) devices,” Br. J. Radiol. **78**, 26–30 (2005).

[14] R. L. Birdwell, P. Bandodkar, and D. M. Ikeda, “Computer-aided detection with screening mammography in a university hospital setting,” Radiology **236**, 451–457 (2005).

[15] M. J. Morton, D. H. Whaley, K. R. Brandt, and K. K. Amrami, “Screening mammograms: Interpretation with computer-aided detection—Prospective evaluation,” Radiology **239**, 375–383 (2006).

[16] J. Wei, H. P. Chan, B. Sahiner, L. M. Hadjiiski, M. A. Helvie, M. A. Roubidoux, C. Zhou, and J. Ge, “Dual system approach to computer-

aided detection of breast masses on mammograms," Med. Phys. **33**, 4157–4168 (2006).

[17] S. Paquerault, N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Improvement of computerized mass detection on mammograms: Fusion of two-view information," Med. Phys. **29**, 238–247 (2002).

[18] B. Zheng *et al.*, "Multiview-based computer-aided detection scheme for breast masses," Med. Phys. **33**, 3135–3143 (2006).

[19] S. van Engeland, S. Timp, and N. Karssemeijer, "Finding corresponding regions of interest in mediolateral oblique and craniocaudal mammographic views," Med. Phys. **33**, 3203–3212 (2006).

[20] B. Liu, C. E. Metz, and Y. Jiang, "An ROC comparison of four methods of combining information from multiple images of the same patient," Med. Phys. **31**, 2552–2563 (2004).

[21] E. A. Krupinski, "Computer-aided detection in clinical environment: benefits and chalenges for radiologists," Radiology **231**, 7–9 (2004).

[22] K. Horsch *et al.*, "Prevalence-modified estimation of computer-determined probabilities of malignancy for CAD," Radiology **223**(P), 388–388 (2003).

[23] B. Zheng, A. Lu, L. A. Hardesty, J. H. Sumkin, C. M. Hakim, A. Ganott, and D. Gur, "A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment," Med. Phys. **33**, 111–117 (2006).

[24] Y. H. Chang, L. A. Hardesty, C. M. Hakim, T. S. Chang, B. Zheng, W. F. Good, and D. Gur, "Knowledge-based computer-aided mass detection on digitized mammograms: a preliminary assessment," Med. Phys. **28**, 455–461 (2001).

[25] G. D. Tourassi, R. Vargas-Voracek, and C. E. Floyd, Jr., "Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information," Med. Phys. **30**, 2123–2139 (2003).

[26] G. D. Tourassi and C. E. Floyd, Jr., "Computer-assisted diagnosis of mammographic masses using an information-theoretic image retrieval scheme with BIRADs-based relevance feedback," Proc. SPIE **5370**, 810–816 (2004).

[27] G. D. Tourassi, B. Harrawood, S. Singh, J. Y. Lo, and C. E. Floyd, Jr., "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms," Med. Phys. **34**, 140–150 (2007).

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[29] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multi-modal image registration by maximization of mutual information," IEEE Trans. Med. Imaging **16**, 187–198 (1997).

[30] G. D. Tourassi and C. E. Floyd, Jr., "Knowledge-based detection of mammographic masses: analysis of the impact of database comprehensiveness," Proc. SPIE **5748**, 399–406 (2005).

[31] M. Heath *et al.*, "Current Status of the digital database for screening mammography," in *Digital Mammography* (Kluwer Academic, Boston, MA, 1998). Available: http://marathon.csee.usf.edu/Mammography/Database.html

[32] D. M. Catarious, A. H. Baydush, and C. E. Floyd, Jr., "A mammographic mass CAD system incorporating features from shape, fractal, and channelized Hotelling observer measurements: Preliminary results," Proc. SPIE **5032**, 111–119 (2003).

[33] D. M. Catarious, A. H. Baydush, and C. E. Floyd, Jr., "Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system," Med. Phys. **31**, 1512–1520 (2004).

[34] N. A. Obuchowski, "Receiver operating characteristic curves and their use in radiology," Radiology **229**, 3–8 (2003).

[35] Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," Radiology **201**, 745–750 (1996).

Computer Aided Detection of Breast Masses in Tomosynthesis Reconstructed Volumes Using Information-Theoretic Similarity Measures

Swatee Singh[1,3], Georgia D. Tourassi[2,3], Amarpreet S. Chawla[1,3], Robert S. Saunders[3], Ehsan Samei[1,2,3], Joseph Y. Lo[1,2,3]

[1]Department of Biomedical Engineering
[2]Duke Medical Physics Graduate Program
[3]Duke Advanced Imaging Laboratories, Department of Radiology
Duke University Medical Center
2424 Erwin Road, Suite 302, Duke University
Durham, NC 27705
E-mail: swatee.singh@duke.edu

ABSTRACT

The purpose of this project is to study two Computer Aided Detection (CADe) systems for breast masses for digital tomosynthesis using reconstructed slices. This study used eighty human subject cases collected as part of on-going clinical trials at Duke University. Raw projections images were used to identify suspicious regions in the algorithm's high sensitivity, low specificity stage using a Difference of Gaussian filter. The filtered images were thresholded to yield initial CADe hits that were then shifted and added to yield a 3D distribution of suspicious regions. The initial system performance was 95% sensitivity at 10 false positives per breast volume. Two CADe systems were developed. In system A, the central slice located at the centroid depth was used to extract a 256x 256 Regions of Interest (ROI) database centered at the lesion coordinates. For system B, 5 slices centered at the lesion coordinates were summed before the extraction of 256x 256 ROIs. To avoid issues associated with feature extraction, selection, and merging, information theory principles were used to reduce false positives for both the systems resulting in a classifier performance of 0.81 and 0.865 Area Under Curve (AUC) with leave-one-case-out sampling. This resulted in an overall system performance of 87% sensitivity with 6.1 FPs/ volume and 85% sensitivity with 3.8 FPs/ volume for systems A and B respectively. This system therefore has the potential to detect breast masses in tomosynthesis data sets.

1. INTRODUCTION

Breast cancer is the second-most deadly type of cancer for women in the United States, second only to lung cancer. The American Cancer Society estimates that 240,510 women will be diagnosed with breast cancer in 2007. They also estimate that breast cancer will kill an estimated 40,910 women in the same year [1]. Survival rates are significantly higher when the cancer is detected at an early stage [2-4]. Several groups have developed CADe algorithms for mammography [5-22]. This study seeks to develop a 3-D Computer Aided Detection (CADe) system for the task of mass detection by using data obtained from a prototype Siemens tomosynthesis system for breast. Despite this new modality's promise to increase sensitivity and reduce unnecessary biopsies, rapid and widespread adoption of tomosynthesis might be impeded by the increase in radiologist reading time per case given the large volume of data generated. Thus, CADe for breast tomosynthesis may be crucial not just to locate overlooked lesions as with mammography, but also to streamline radiologist workflow when interpreting such a large volume of data. As current investigators in CT colonography have suggested, CADe can potentially ease radiologist workflow when working with large 3D data sets [23].

Medical Imaging 2008: Computer-Aided Diagnosis, edited by Maryellen L. Giger, Nico Karssemeijer
Proc. of SPIE Vol. 6915, 691505, (2008) · 1605-7422/08/$18 · doi: 10.1117/12.772978

2. METHODS AND MATERIALS

2.1 Database

A prototype breast tomosynthesis system by Siemens Medical Solutions was developed to acquire 25 projection images over a 50-degree angular range in approximately 13 seconds. Resultant projection images from this system are of high resolution (85 micron pixel size), and are acquired at the rate of 2 images/second frames. So far, over 235 human subjects have been recruited at the Duke University Medical Center. Bilateral MLO views were acquired in screening cases, while bilateral MLO and CC views were acquired for diagnostic and biopsy cases. A single MQSA breast imaging radiologist with 15 years experience interpreted these cases in separate and blinded readings. The gold standard was established from information available from all modalities for a patient. For this study, we used data from eighty patients of which 20% contained a lesion in at least one view, while the other 80% were completely normal cases.

We wished to draw upon our existing mammography-based CADe techniques, and by working with projection images - which are similar to low dose mammogram images - we were able to achieve that objective to identify an initial set of suspicious regions. While our algorithm's initial stage is reconstruction algorithm independent, the second stage uses reconstructed slices for FP reduction. Other groups have implemented completely reconstruction independent CADe [24].

2.2 Experimental Design

Our CADe algorithm can be divided into two stages – (1) the high sensitivity, low specificity stage, (2) False Positive (FP) stage. Figures 1 and 2 graphically depict these two stages.

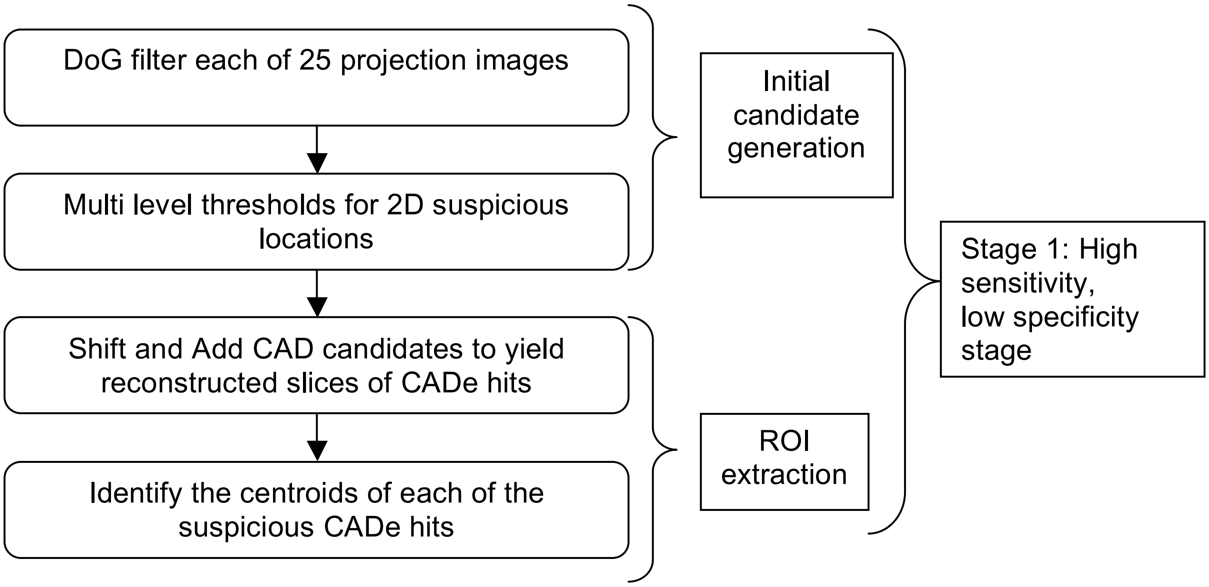

DoG filter each of 25 projection images
Multi level thresholds for 2D suspicious locations
Shift and Add CAD candidates to yield reconstructed slices of CADe hits
Identify the centroids of each of the suspicious CADe hits
Initial candidate generation
ROI extraction
Stage 1: High sensitivity, low specificity stage

Figure 1: The first stage of the algorithm adopted in this study for our CADe system

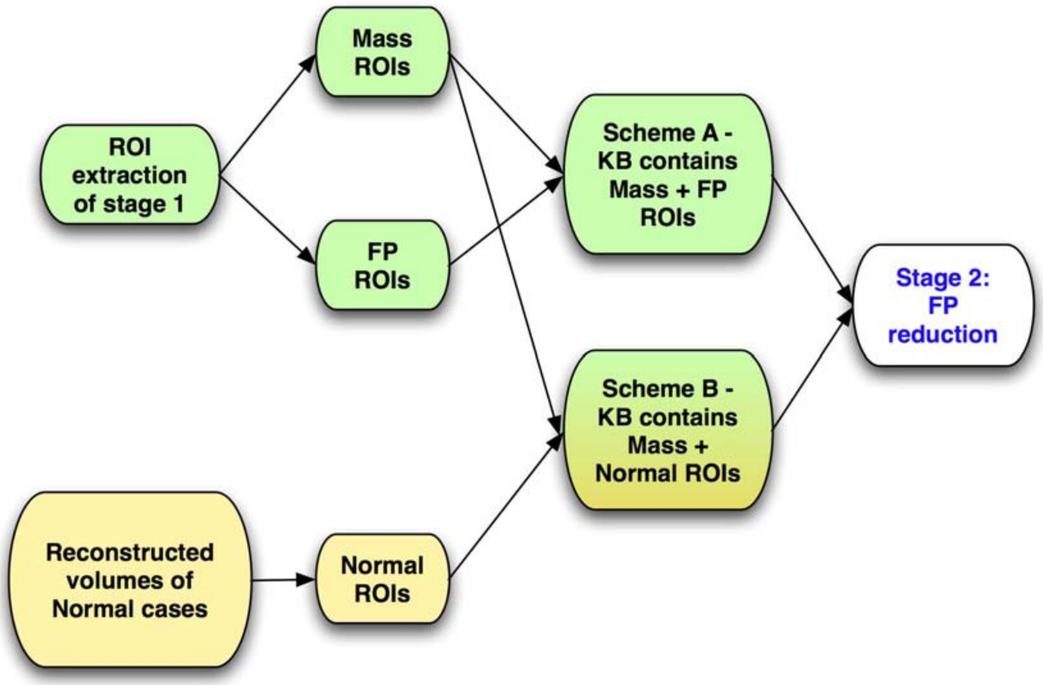

Mass ROIs
ROI extraction of stage 1
FP ROIs
Scheme A - KB contains Mass + FP ROIs
Stage 2: FP reduction
Scheme B - KB contains Mass + Normal ROIs
Reconstructed volumes of Normal cases
Normal ROIs

Figure 2: The second stage of the algorithm adopted in this study for our CADe system

2.2.1 High-sensitivity, low-specificity stage

Projection images contain greater intensity variation across lesion and non- lesion locations when compared to reconstructed slices. Hence, to identify initial suspicious locations we chose to work with projection images rather than reconstructed slices as they are likely to retain the maximum possible information about the lesion. Other groups have tried a similar approach [24, 25]. Potential suspicious locations were identified using a Difference of Gaussian (DoG) filter on each of the projection images. These segmented lesion candidates in every projection image were shifted and added using the acquisition angle and known geometry. A typical resulting 3-D volume of CADe suspicious locations is shown in figure 3. The CADe suspicious locations were connected in 3D space using a 3x3x3 connectivity rule to determine location and shape of the object. With shift and add, significant out of plane blur is observed. However, the true object lies in the plane with the least area of the classic starburst shape obtained via the shift and add algorithm. A graphical representation of this stage of the algorithm is displayed in figure 1.

Once the first stage of the CADe algorithm identified initial candidates for mass detection by giving us the X, Y and Z location of the centroid of the volume of interest (VOI), we extracted ROIs from the reconstructed breast slice images that the radiologists look at. These breast volume reconstructions were obtained by filtered backprojection reconstruction [26]. The first stage of the CADe algorithm yielded a 95% sensitivity at 10 false positives per breast volume for both systems.

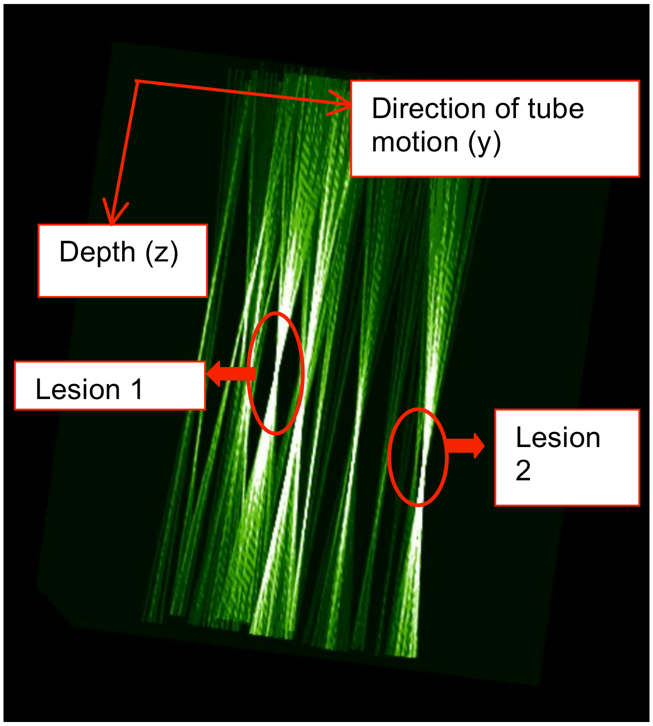


Figure 3: Maximum intensity projection image of subject 33's CAD suspicious locations. The mass comes in focus where the starburst shapes in the YZ plane have the least area. Significant out of plane blurring is observed with shift and add reconstruction of the CAD suspicious locations. Both lesions were detected by this stage of the CAD algorithm.

2.2.1 False Positive Reduction

We used a second stage false positive reduction algorithm that relies on information theoretic principles to assess image similarity. We investigated two false positive reduction schemes and they are graphically shown in figure 2. For system A, we obtained 256x 256 ROIs centered at the central slice containing the suspicious location given by the first stage of the algorithm. For system B, we used the same locations, however we summed 5 reconstructed slices centered at the location. The motivation behind this was that lesions typically span multiple reconstructed slices and we wished to investigate whether giving more 'signal' to our false positive reduction scheme resulted in an improvement in performance.

In our algorithm's false positive reduction stage, the query ROI is compared to a knowledge database of ROIs with known ground truth. Similar cases based on similarity metrics such as mutual information are retrieved from the knowledge database. A decision is formulated regarding the query region using the retrieved similar cases. If the query region depicts a mass, then the calculated decision index is higher than if it contains normal breast tissue [27, 28]. ROIs obtained from both system A and B were used separately as knowledge database and tested using a leave-one-case-out sampling scheme.

For this experiment, we measured mutual information as a similarity metric. Mutual information (MI) is a measure of the information one random information contains about the other. Hence knowledge of the first random variable reduces the uncertainity in predicting the value of the second random variable. It is given by the following equation [29]:

$$MI(X,Y) = \sum_x \sum_y P_{XY}(X,Y)\log_2\left(\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right) \qquad (1)$$

where *X* and *Y* are two random variables, $p(x,y)$ is their joint probability mass function because this is a discrete rather than continuous random variable and $p(x)$ and $p(y)$ are the marginal probability mass functions of *X* and *Y*. Further details about these measures and their specific application have been previously published [27, 28].

However, we still need to obtain a CADe score for a given ROI of unknown pathology. This is done via the adoption of a decision index. Given a query tomosynthesis ROI *Qi* , a decision index *D*(*Qi*) was calculated by our algorithm as the difference of two terms. Assuming that the knowledge database contains *k* mass cases and *l* normal cases. The first term of the decision index *D*(*Qi*) measures the average MI between the query ROI and its *k* best mass matches M_j . Similarly, the second term measures the average MI between the query ROI and its *l* best normal N_j matches,

$$D(Q_i) = \frac{1}{k}\sum_{j=1}^{k} MI(Q_i, M_j) - \frac{1}{l}\sum_{j=1}^{l} MI(Q_i, N_j) \qquad (2)$$

Theoretically, a query ROI depicting a mass should have a higher *D*(*Qi*). For this study eighty cases were used. This algorithm's initial high sensitivity, low specificity stage yielded ROIs that were extracted from the middle projection. Hence, false positive reduction was done using only ROIs obtained from the middle projection image of each scan. Results were reported as Receiver Operating Characteristic (ROC) Area Under Curve (AUC) by applying a leave-one-out cross validation scheme on all available ROIs.

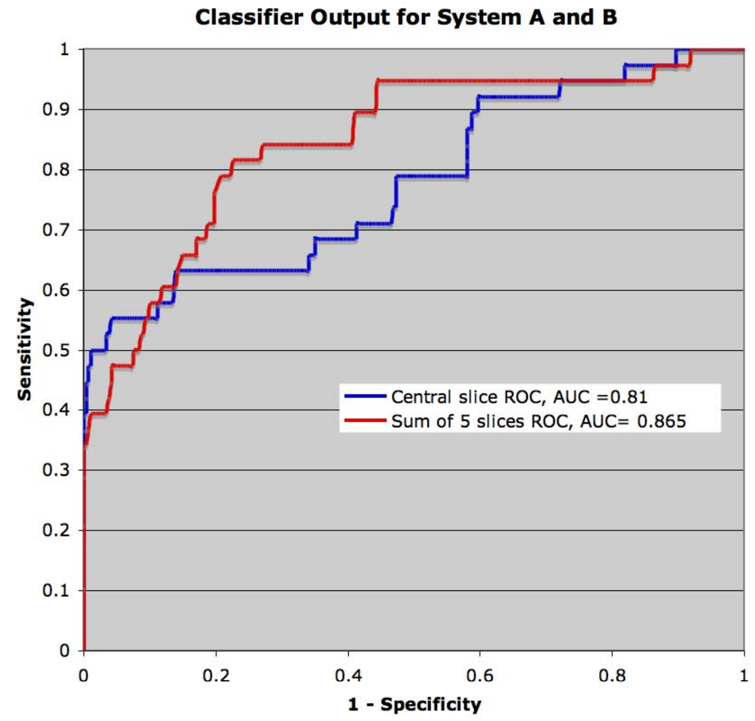
3. RESULTS

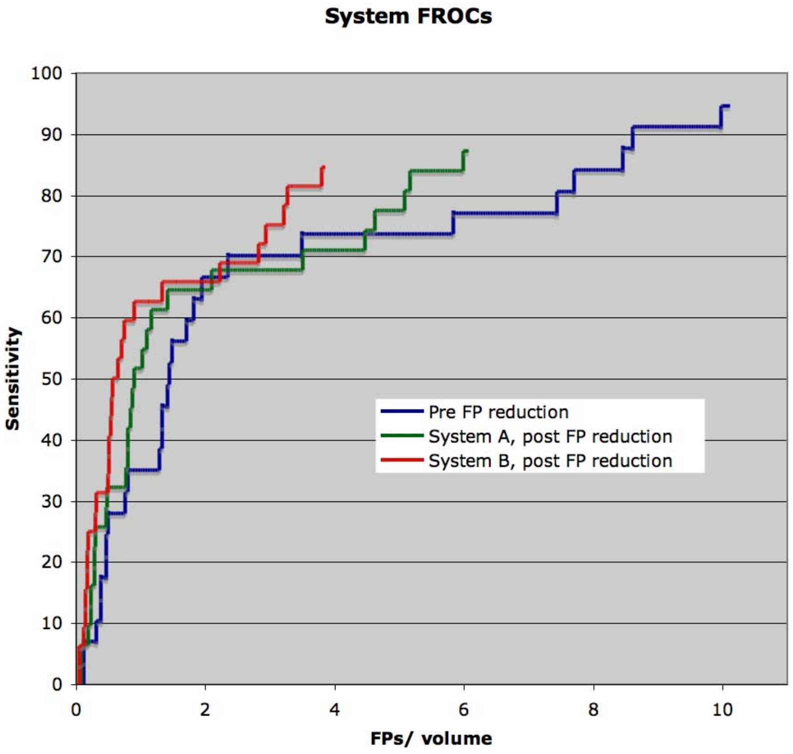
We used Mutual Information (MI) as a similarity metric for the two false positive reduction schemes. The results for both the systems are tabulated in table 1 below.

	Classifier AUC	**Classier partial AUC (TPF >=0.90)**
System A	0.81 +/- 0.04	0.10
System B	0.865 +/- 0.04	0.30

Table 1: Performance of the false positive reduction stage of the CAD algorithm for the two systems A and B using mutual information as the similarity metric.

While the performances in the important high-sensitivity range of the classifier output were both good, system B showed an advantage. The two classifier ROC curves are shown in figure 4. When the two false positive reduction schemes were applied to the initial suspicious regions at the threshold obtained for 92% and 89% classifier sensitivities, the final result over this data set was 87% with 6.1 FPs/ volume and 85% sensitivity with 3.8 FPs/ volume for systems A and B respectively. The final FROC curves for both systems, pre and post false positive reduction stage are shown below in figure 5.


Classifier Output for System A and B
Sensitivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
1 - Specificity
Central slice ROC, AUC =0.81
Sum of 5 slices ROC, AUC= 0.865


System FROCs
Sensitivity
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
FPs/ volume
Pre FP reduction
System A, post FP reduction
System B, post FP reduction

4. CONCLUSIONS

We have demonstrated viability of a promising CADe algorithm using models with the extremely low-dose tomosynthesis projection slices and information theoretic principles for false positive reduction. Future work will expand the use of information theory principles to be fully 3D for reduction of false positives. Concurrently, we will continue to work towards increasing the size of our database, and will explore direct optimization of the CAD techniques for tomosynthesis as well.

ACKNOWLEDGMENTS

This work was supported in part by US Army Breast Cancer Research Program W81XWH-05-1-0293, NIH/NCI R01 CA101911 and R01 CA112437, and Siemens Medical Solutions. We would like to thank Thomas Mertelmeier PhD and Jasmina Ludwig of Siemens Medical Solutions for development of the FBP reconstruction software and helpful suggestions. Also, the authors thank all the radiologists at Duke University Medical Center who have participated in this study. The authors would like to extend special thanks to Brian Harrawood for his support in scientific programming for this study.

REFERENCES

[1] ACS, "American Cancer Society: Cancer Facts and Figures 2007-2008. Atlanta, Ga: American Cancer Society 2007.," (2007).

[2] Anttinen, I., Pamilo, M., Soiva, M. and Roiha, M., "Double reading of mammography screening films: one radiologist or two?," Clinical Radiology 48, 414-421 (1993).

[3] Hendee, W. R., Beam, C. and Hendrick, E., "Proposition: all mammograms should be double-read," Medical Physics 26, 115-118 (1999).

[4] Thurfjell, E. L., Lernevall, K. A. and Taube, A. A. S., "Benefit of independent double reading in a population-based mammography screening program," Radiology 191, 241-244 (1994).

[5] Wu, Y.-T., Wei, J., Hadjiiski, L. M., Sahiner, B., Zhou, C., Ge, J., Shi, J., Zhang, Y. and Chan, H.-P., "Bilateral analysis based false positive reduction for computer-aided mass detection," Medical Physics 34(8), 3334-3344 (2007).

[6] Wei, J., Chan, H.-P., Sahiner, B., Hadjiiski, L. M., Helvie, M. A., Roubidoux, M. A., Zhou, C. and Ge, J., "Dual system approach to computer-aided detection of breast masses on mammograms," Medical Physics 33(11), 4157-4168 (2006).

[7] Sahiner, B., Chan, H.-P., Hadjiiski, L. M., Helvie, M. A., Paramagul, C., Ge, J., Wei, J. and Zhou, C., "Joint two-view information for computerized detection of microcalcifications on mammograms," Medical Physics 33(7), 2574-2585 (2006).

[8] Wei, J., Sahiner, B., Hadjiiski, L. M., Chan, H.-P., Petrick, N., Helvie, M. A., Roubidoux, M. A., Ge, J. and Zhou, C., "Computer-aided detection of breast masses on full field digital mammograms," Medical Physics 32(9), 2827-2838 (2005).

[9] Paquerault, S., Petrick, N., Chan, H. P., Sahiner, B. and Helvie, M. A., "Improvement of computerized mass detection on mammograms: Fusion of two-view information," Medical Physics 29(2), 238-247 (2002).

[10] Qian, W., Li, L. and Clarke, L. P., "Image feature extraction for mass detection in digital mammography: Influence of wavelet analysis," Medical Physics 26(3), 402-408 (1999).

[11] Yu, S. and Guan, L., "A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films," IEEE Trans Med Imag 19(2), 115-126 (2000).

[12] Schmidt, F., Sorantin, E., Szepesvari, C., Graif, E., Becker, M., Mayer, H. and Hartwagner, K., "An automatic method for the identification and interpretation of clustered microcalcifications in mammograms," Phys Med Biol 44(5), 1231-1243 (1999).

[13] Huo, Z., Giger, M. L., Vyborny, C. J., Wolverton, D. E. and Metz, C. E., "Computerized classification of benign and malignant masses on digitized mammograms: a study of robustness," Acad Radiol 7(12), 1077-1084 (2000).

[14] Catarious, D. M., Baydush, A. H., Abbey, C. K. and Floyd, C. E., Jr, "A Mammographic mass CAD system incorporating features from shape, fractal, and channelized Hotelling observer measurements: preliminary results," in Proceedings of Proc. SPIE Int. Soc. Opt. Eng., 111 (2003).

[15] Catarious, D. M., Jr., Baydush, A. H. and Floyd, C. E., Jr., "Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system," Med Phys 31(6), 1512-1520 (2004).

[16] Sampat, M. P., Markey, M. K. and Bovik, A. C., "Computer-aided detection and diagnosis in mammography," in [Handbook of Image and Video Processing], edited by A. C. Bovik, pp 1195-1217, Academic Press, (2005).

[17] Zheng, B., Chang, Y. H., Wang, X. H., Good, W. F. and Gur, D., "Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm," Acad Radiol 6(6), 327-332 (1999).

[18] Chan, H. P., Sahiner, B., Lam, K. L., Petrick, N., Helvie, M. A., Goodsitt, M. M. and Adler, D. D., "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces," Medical Physics 25(10), 2007-2019 (1998).

[19] Gavrielides, M. A., Lo, J. Y., Vargas-Voracek, R. and Floyd, C. E., Jr, "Segmentation of suspicious clustered microcalcifications in mammograms," Medical Physics 27(1), 13-22 (2000).

[20] Ge, J., Sahiner, B., Hadjiiski, L. M., Chan, H.-P., Wei, J., Helvie, M. A. and Zhou, C., "Computer aided detection of clusters of microcalcifications on full field digital mammograms," Medical Physics 33(8), 2975-2988 (2006).

[21] Wei, J., Sahiner, B., Hadjiiski, L. M., Chan, H.-P., Petrick, N., Helvie, M. A., Roubidoux, M. A., Ge, J. and Zhou, C., "Computer-aided detection of breast masses on full field digital mammograms," Medical Physics 32(9), 2827 (2005).

[22] Li, L., Zheng, Y., Zheng, L. and Clark, R. A., "False-positive reduction in CAD mass detection using a competitive classification strategy," Medical Physics 28(2), 250-258 (2001).

[23] Abraham, H. D. and Hiro, Y., "Virtual colonoscopy: past, present,and future," Radiologic clinics of North America 41(2), 377-393 (2003).

[24] Reiser, I. S., Sidky, E. Y., Giger, M. L., Nishikawa, R. M., Rafferty, E. A., Kopans, D. B., Moore, R. and Wu, T., "A reconstruction-independent method for computerized mass detection in digital tomosynthesis images of the breast," in Proceedings of Medical Imaging 2004: Image Processing, 833-838 (2004).

[25] Chan, H.-P., Wei, J., Zhang, Y., Moore, R. H., Kopans, D. B., Hadjiiski, L., Sahiner, B., Roubidoux, M. A. and Helvie, M. A., "Computer-aided detection of masses in digital tomosynthesis mammography: combination of 3D and 2D detection information," in Proceedings of Medical Imaging 2007: Computer-Aided Diagnosis, 651416-651416 (2007).

[26] Mertelmeier, T., Orman, J., Haerer, W. and Dudam, M. K., "Optimizing filtered backprojection reconstruction for a breast tomosynthesis prototype device," in Proceedings of Medical Imaging 2006: Physics of Medical Imaging, 61420F-61412 (2006).

[27] Tourassi, G. D., Harrawood, B., Singh, S. and Lo, J. Y., "Information-theoretic CAD system in mammography: Entropy-based indexing for computational efficiency and robust performance," Medical Physics 34(8), 3193-3204 (2007).

[28] Tourassi, G. D., Harrawood, B., Singh, S., Lo, J. Y. and Floyd, C. E., Jr., "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms," Medical Physics 34(1), 140-150 (2007).

[29] Cover, T. and Thomas, J., [Elements of information theory], Wiley-Interscience, (1991).

Effect of ROI Size on the Performance of an Information-Theoretic CAD System in Mammography: Multi-size Fusion Analysis

Robert C. Ike III[1], Swatee Singh[1], Brian Harrawood[1], Georgia D. Tourassi[1]

[1] Digital Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, NC 27705

ABSTRACT

Featureless, knowledge-based CAD systems are an attractive alternative to feature-based CAD because they require no to minimal image preprocessing. Such systems compare images directly using the raw image pixel values rather than relying on low-level image features. Specifically, information-theoretic (IT) measures such as mutual information (MI) have been shown to be an effective, featureless, similarity measure for image comparisons. MI captures the statistical relationship between the gray level values of corresponding image pixels. In a CAD system developed at our laboratory, the above concept has been applied for location-specific detection of mammographic masses. The system is designed to operate on a fixed size region of interest (ROI) extracted around a suspicious mammographic location. Since mass sizes vary substantially, there is a potential drawback. When two ROIs are compared, it is unclear how much the parenchymal background contributes in the calculated MI. This uncertainty could deteriorate CAD performance in the extreme cases, namely when a small mass is present in the ROI or when a large mass extends beyond the fixed size ROI. The present study evaluates the effect of ROI size on the overall CAD performance and proposes multisize analysis for possible improvement. Based on two datasets of ROIs extracted from DDSM mammograms, there was a statistically significant decline of the CAD performance as the ROI size increased. The best size ranged between 512x512 and 256x256 pixels. Multisize fusion analysis using a linear model achieved further improvement in CAD performance for both datasets.

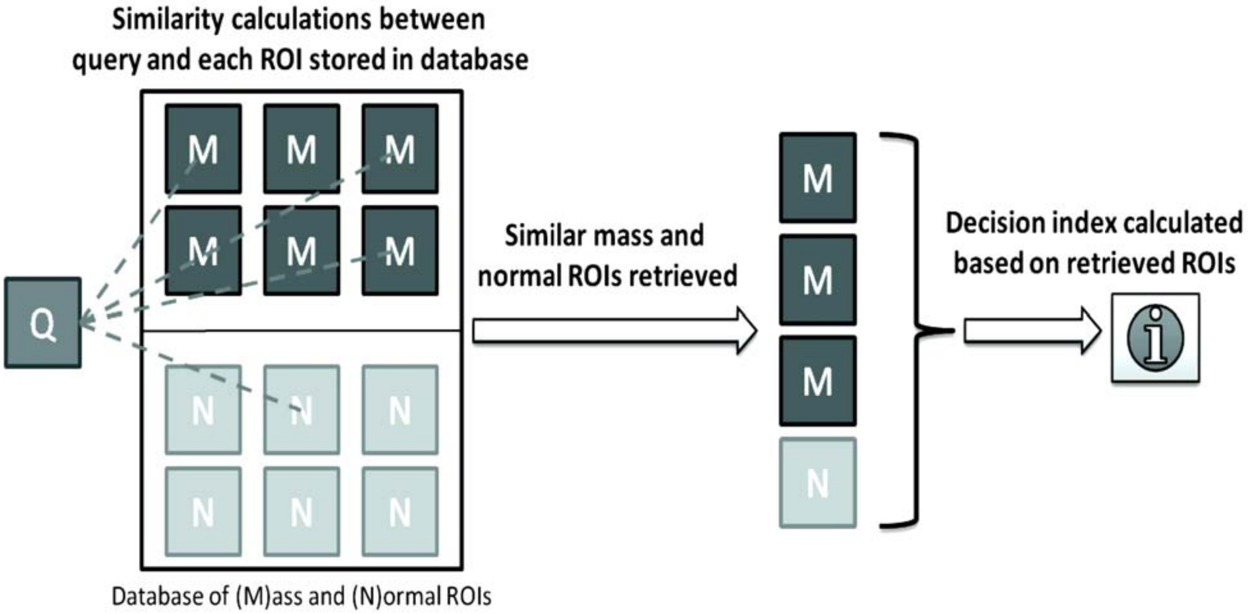
Keywords: classification and classifier design, mammography, detection

1. INTRODUCTION

The task for a radiologist of reading and properly identifying all lesions in a mammogram accurately and efficiently is a difficult one. It is hoped that an effective CAD system will provide the radiologist with a reliable second opinion, leading to high sensitivity. While clinical studies have shown that the addition of a CAD system can increase the sensitivity from 16.1%[1] up to 19.5%[2], a major complaint is the lack of specificity, or higher than desired false positive rate[3].

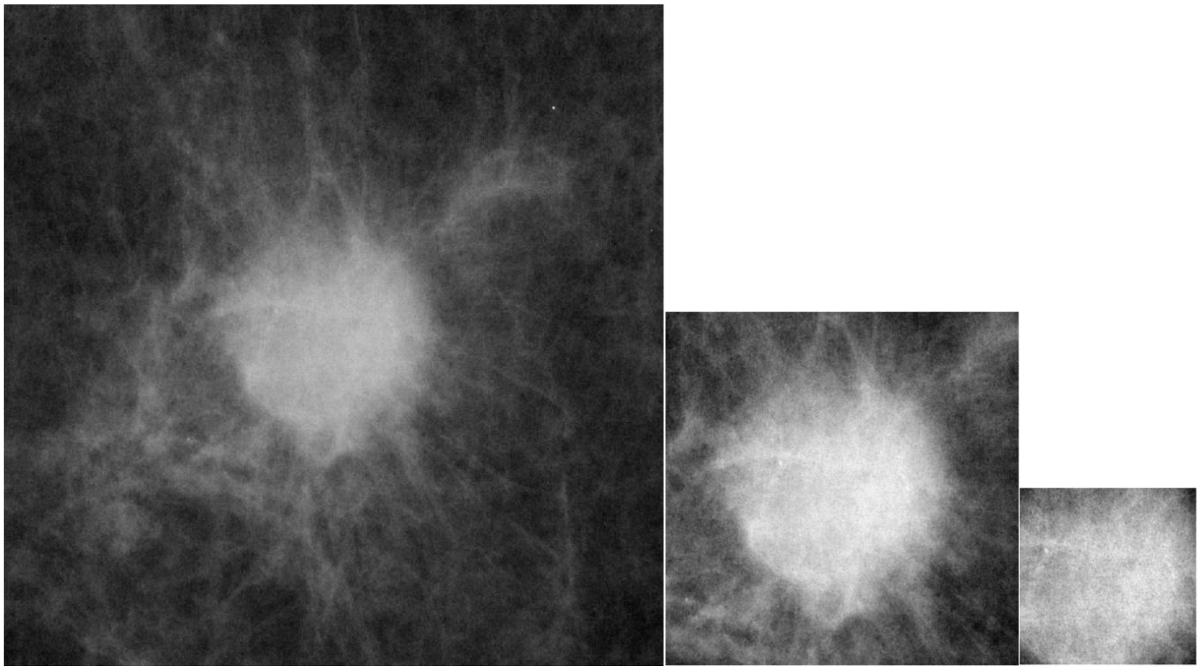
In earlier studies, we introduced a featureless, information-theoretic CAD (IT-CAD) system for the detection of masses in regions-of-interest (ROIs) deemed suspicious during screening[4,5,6]. This was accomplished by extracting 512x512 pixel ROIs around the suspicious image locations and comparing them to a knowledge base of other 512x512 pixel ROIs with known ground truth. However, rather than comparing features such as shape, size, and other physical features, as many CAD systems do, our system relies on information theoretic principles to determine similarity. In our IT-CAD system, mutual information (MI) measures are used to assess the similarity between ROIs. However, this fixed size ROI approach may be inadequate. Masses can vary in size from thousands of pixels to just tens of pixels. This means that for small masses, background parenchyma would comprise most of the information in the ROI, while some of the larger masses may not be fully contained within the ROI. It would seem that a variable ROI size scaled to match the size of the mass would be better suited to capturing all the relevant information.

Medical Imaging 2008: Computer-Aided Diagnosis, edited by Maryellen L. Giger, Nico Karssemeijer
Proc. of SPIE Vol. 6915, 691527, (2008) · 1605-7422/08/$18 · doi: 10.1117/12.772678


Similarity calculations between query and each ROI stored in database
Q
M
N
Database of (M)ass and (N)ormal ROIs
Similar mass and normal ROIs retrieved
Decision index calculated based on retrieved ROIs
i

$$I(Q;T) = \sum_q \sum_t P_{QT}(q,t) log_2 \frac{P_{QT}(q,t)}{P_Q(q)P_T(t)}$$

$$D(Q) = \frac{1}{K}\sum_{i=1}^{K} MI(Q,M_i) - \frac{1}{K}\sum_{i=1}^{K} MI(Q,N_i)$$



2.3 Experimental Design

To study the impact of ROI size on IT-CAD performance, two experiments were conducted.

Experiment 1: Fixed Size

At each ROI size, the first set of 1,557 ROIs was used in a leave-one-out manner. In other words, each ROI was excluded once to serve as a query while the remaining served as the knowledge database.

Then, dataset 1 was used as the knowledge database and dataset 2 was used as the query database. This experiment was designed to determine the effect of ROI size on the IT-CAD performance when applied for false positive reduction.

Experiment 2: Multi-Size Fusion

The IT-CAD outputs for the 3 respective sizes were merged with a linear discriminant (LDA) decision model to assess whether multisize fusion improves CAD performance. The same leave-one-out sampling scheme was applied for dataset 1 to determine the performance of the multifusion scheme for discriminating masses from normal ROIs. Subsequently, dataset 1 was used to train LDA and then test it on dataset 2 to determine the performance of the multifusion scheme for discriminating masses from false positive ROIs.

2.4 Performance Evaluation

To evaluate performance, a receiver operating characteristics (ROC) analysis[10] was performed for each experiment. The ROCKIT software developed by Metz et al (available at www.radiology.uchicago.edu/krl/toppage 11.htm) was applied with the IT-CAD decision index as the ROC decision variable.

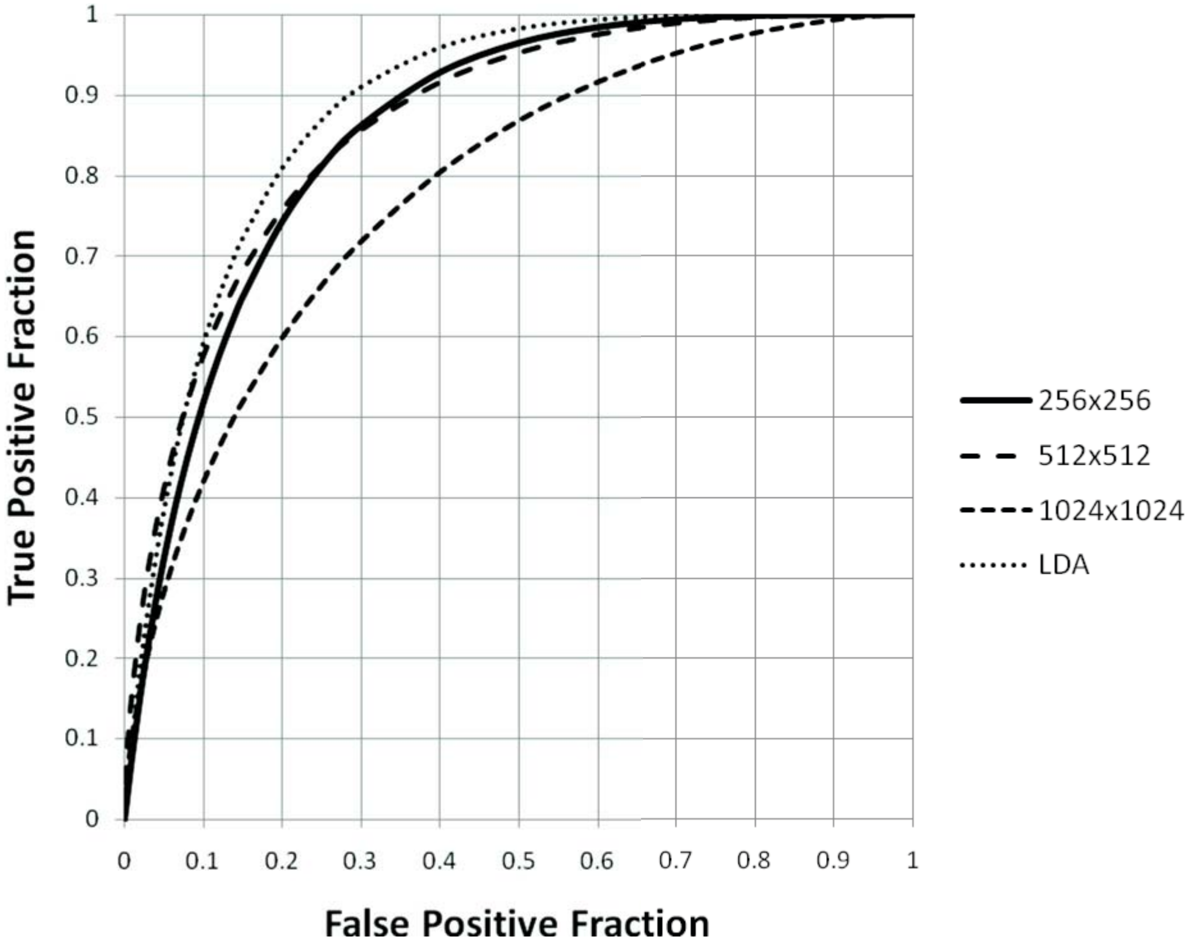
LDA was also applied using R software[11,12], a GNU project developed by the R Foundation for Statistical Computing (available at www.R-project.org).

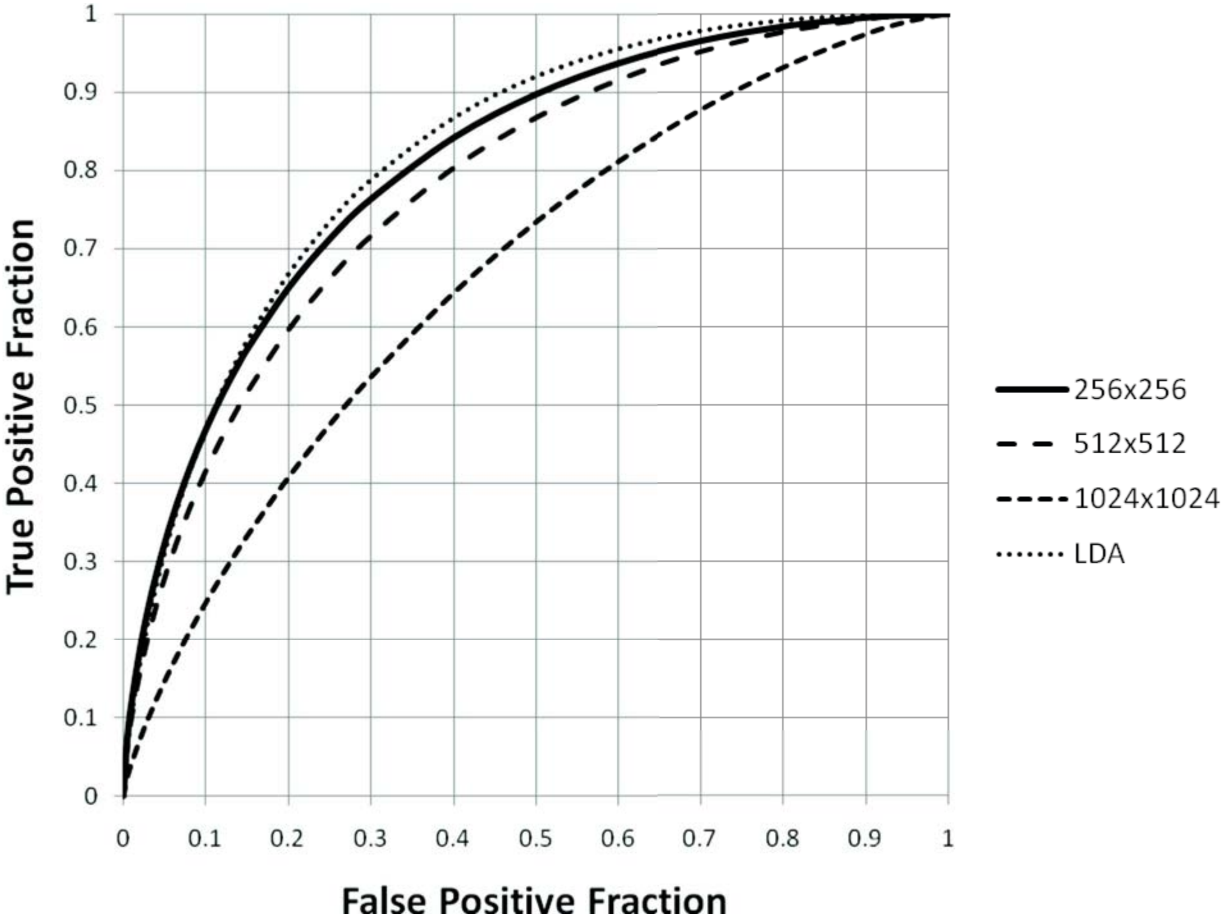
3. RESULTS

Table 1 summarizes the performance of the first experiment. ROC performance varies depending on the ROI size for both datasets. The largest size (1024x1024) provides the worst detection performance in Dataset 1. The performance decline was statistically significant. The best performance was observed for the middle size (512x512) in dataset 1, with no statistically significant advantage (two-tailed p-value=0.35) over the smallest size (256x256). These trends are likely due to many of the masses being "lost" in the vast background of the 1024x1024 ROIs, and a few masses not being entirely included in the 256x256 ROIs.

Table 1: ROC performance of the IT-CAD system depending on the ROI size

ROI SIZE	Az (Masses vs. Normals)	Az (Masses vs. FPS)
256 x 256 pixels	0.857±0.010	0.814±0.025
512 x 512 pixels	0.863±0.009	0.787±0.026
1024 x 1024 pixels	0.777±0.011	0.656±0.032


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
True Positive Fraction
False Positive Fraction
256x256
512x512
1024x1024
LDA



True Positive Fraction
False Positive Fraction
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
256x256
512x512
1024x1024
LDA

4. CONCLUSION

While choosing a single ROI size for our featureless CAD system may provide reasonable overall performance, this approach is suboptimal with masses that are too small or too large due to incomplete inclusion of the candidate mass or inclusion of excessive background. Our results show that there was a statistically significant decrease in our CAD performance when using the large-size ROIs.

As a first step, this study demonstrates that improvements can be achieved with a multi-size fusion approach, as there was a small yet consistent improvement observed with multi-size linear fusion. However, implementing a custom ROI size knowledge database maybe the best strategy to achieve robust detection performance for all mass sizes. Experiments with advanced artificial intelligence fusion techniques are currently in progress.

5. ACKNOWLEDGEMENTS

This work was partially supported by grant R01 CA101911 from the National Cancer Institute.

6. REFERENCES

1. Cupples T E, Cunningham J E, Reynolds J C. Impact of Computer-Aided Detection in a Regional Screening Mammography Program, American Journal of Radiology 2005; 185: 944-950.
2. Freer T W, Ulissey M J. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center, Radiology 2001; 220: 781-786.
3. Taylor P, Given-Wilson R M. Evaluation of computer-aided detection (CAD) devices, British Journal of Radiology 2005; 78: 26-30.
4. Tourassi G D, Vargas-Voracek R, Floyd, Jr. C E. Computer-Assisted Detection of Mammographic Masses: A Template Matching Scheme based on Mutual Information, Medical Physics 2003; 30: 2123-2139.
5. Tourassi G D, Harrawood B, Singh S, Lo J Y, Floyd, Jr. C E. Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms, Medical Physics 2007; 34: 140-150.
6. Tourassi G D, Harrawood B, Singh S, Lo J Y. Information-theoretic CAD system in mammography: Entropy-based indexing for computational efficiency and robust performance, Medical Physics 2007; 34: 3193-3204.
7. Heath M, *et al.*, Current status of the digital database for screening mammography, Digital mammography (Kluwer, Dordrecht, 1998). Available: http://marathon.csee.usf.edu/Mammography/Database.html
8. Catarious D M, Baydush A H, Floyd, Jr. C E. A mammographic mass CAD system incorporating features from shape, fractal, and channelized Hotelling observer measurements: Preliminary results, Proc. SPIE 2003; 5032: 111-119.
9. Catarious D M, Baydush A H, Floyd, Jr. C E. Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system, Medical Physics 2004; 31 (6): 1512-1520.
10. Obuchowski N A. Receiver Operating Characteristic curves and their use in radiology, Radiology 2003; 229: 3-8.
11. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2003.
12. The R project for statistical computing (http://www.R-project.org)

Optimized Acquisition Scheme for Multi-projection Correlation Imaging of Breast Cancer

Amarpreet S. Chawla, Ehsan Samei, Robert S. Saunders, Joseph Y. Lo, and Swatee Singh
Duke Advanced Imaging Laboratories (DAILabs), Duke University

ABSTRACT

We are reporting the optimized acquisition scheme of multi-projection breast Correlation Imaging (CI) technique, which was pioneered in our lab at Duke University. CI is similar to tomosynthesis in its image acquisition scheme. However, instead of analyzing the reconstructed images, the projection images are directly analyzed for pathology. Earlier, we presented an optimized data acquisition scheme for CI using mathematical observer model. In this article, we are presenting a Computer Aided Detection (CADe)-based optimization methodology. Towards that end, images from 106 subjects recruited for an ongoing clinical trial for tomosynthesis were employed. For each patient, 25 angular projections of each breast were acquired. Projection images were supplemented with a simulated 3 mm 3D lesion. Each projection was first processed by a traditional CADe algorithm at high sensitivity, followed by a reduction of false positives by combining geometrical correlation information available from the multiple images. Performance of the CI system was determined in terms of free-response receiver operating characteristics (FROC) curves and the area under ROC curves. For optimization, the components of acquisition such as the number of projections, and their angular span were systematically changed to investigate which one of the many possible combinations maximized the sensitivity and specificity. Results indicated that the performance of the CI system may be maximized with 7-11 projections spanning an angular arc of 44.8^{o}, confirming our earlier findings using observer models. These results indicate that an optimized CI system may potentially be an important diagnostic tool for improved breast cancer detection.

KEYWORDS: Multi-projection Imaging, Correlation Imaging, Breast Tomosynthesis, FROC, CADe.

INTRODUCTION

Medical imaging is fast advancing towards multi-projection imaging. In this technique, multiple images of the same patient are acquired from slightly different angles. The correlative information between different angular projections is then processed to extract knowledge about the presence as well as the morphology of a potential pathology in the patient. Multi-projection imaging technique thus builds on the advantages of standard projection techniques and combines it with the proven benefits of fusing information from multiple images, and has been proven to potentially improve the accuracy of cancer detection.[1, 2] In digital radiographic imaging, this imaging scheme can take the form of Tomosynthesis,[3] Correlation Imaging,[2] or Stereoscopic Imaging.[4]

While multi-projection imaging technique has notable potentials, in developing such a technique, an important consideration is its data acquisition scheme. Multiple aspects of data acquisition can influence the performance of this technique. The diagnostic outcome is a function of the number of images acquired, the total angular span of these acquisitions and the clinical dose at which these images are acquired. An optimum image acquisition scheme of an imaging system is a specific combination of those various components of acquisition that maximizes the available diagnostic information. This critical aspect of acquisition was studied in this work.

We have previously presented a mathematical observer model-based methodology to optimize the geometry of data acquisition scheme for Correlation Imaging.[5] The purpose of this study is to optimize the geometry using a CADe-based technique and to substantiate the results with those obtained from the mathematical observer model. As a key step towards that goal, a new CADe system for CI was developed. This CADe technique was based on a CADe processor reported earlier.[6] To optimize the geometry of acquisitions, the acquisition parameters were systematically changed and the CAD-based performance measured for different settings of those parameters. An optimized acquisition scheme was defined as the one that generated the best CADe

Medical Imaging 2008: Computer-Aided Diagnosis, edited by Maryellen L. Giger, Nico Karssemeijer
Proc. of SPIE Vol. 6915, 691528, (2008) · 1605-7422/08/$18 · doi: 10.1117/12.773174

performance. The optimization framework presented here is generic in nature and may be used to optimize any multi-acquisition scheme, including tomosynthesis.

MATERIALS AND METHODS

A. Image Database

The study employed a database of image sets from 106 subjects recruited for our ongoing tomosynthesis clinical trial. Each image set consisted of 25 images of a subject acquired from different but fixed angular positions equally spaced over a ~50^o arc by a prototype clinical multi-projection system, Siemens' Mammomat Novation TOMO (Fig. 1). The images were acquired at kVps ranging between 28 and 30, while the total dose delivered to the patient was equivalent to that delivered in a standard two-view screening procedure. All cases were judged to be normal (without any lesions).

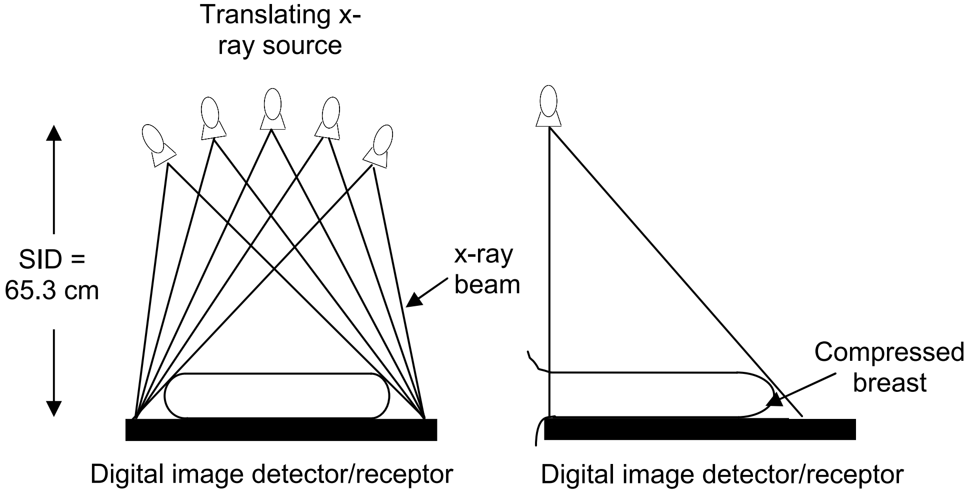


Fig. 1: Schematic of acquisition for multi-projection breast Correlation Imaging (CI). Front view (left); side view (right)

A 3 mm 3D lesion was simulated [5] and its projection embedded into 53 out of the available 106 cases in the database, creating two image datasets, one with lesion-absent and the other with lesion-present. The contrast of the lesion was modified based on the acquisition kVp, target/filter combination, breast thickness, anode type, and scatter fraction. A previously reported routine was used for this purpose.[7]

B. CADe processor

A computer-aided detection (CADe) processor was developed to investigate the performance of CI in terms of detectability of the embedded simulated mass.[6] Specifically, the projection images were first filtered using an adaptive Gaussian gradient filter, which results in a blurry estimate of the anatomical background and highlights suspicious abnormalities in the images. Following filtration, the images were segmented using a 3D segmentation technique to enhance the suspicious regions. The segmentation was optimized to highlight structures with sizes similar to the expected 3 mm embedded lesions. These highlighted regions were then grown using a grayscale region growing technique and finally thresholded to remove smaller unwanted structures. Next, the segmented images were processed with a false positive reduction step that reduced the segmented structures in the images based on their morphological features. Specifically, nine morphological features were used and combined using a genetic algorithm-based decision fusion scheme that determined optimum feature thresholds that were used to determine if a structure was a potential candidate for a lesion.

The segmented images were then processed by a shift and add reconstruction technique to generate a CADe-enhanced volume within which a potential lesion was segmented. This stack of slices were then collapsed into a single 2D image that brings into focus the most suspected regions, while the regions with a less likelihood for the presence of a lesion were blurred out. A thresholding mechanism was applied to pick the region with the suspected pathology.

C. Optimization of Data Acquisition

To optimize the acquisition scheme, the components of acquisition such as the number of projections, and their angular span were systematically changed to investigate which one of the many possible combinations yield the highest diagnostic performance.

The diagnostic performance was measured in terms of two performance indices. First, the ratio of True Positives/(True Positives + False Positives), termed Positive Predictive Index (PPI), was used as a measure of the true positive locations as a fraction of the total number of identified locations *per image set*. These values were obtained for different combinations of the number of projection and angular range and then averaged across all the cases for each combination of the acquisition setting.

The second performance index was the area under the ROC curve (AUC). The AUC was computed using the datasets with and without the embedded lesion. Each case in the two datasets was processed with the CADe processor to obtain 2D contour maps. Next, instead of analyzing the final collapsed 2D contour map for the number of false-positives per case, only the likelihood of the presence of the embedded lesion was investigated. Specifically, a correlation matching of the expected signal with the signal-present and signal-absent 2D contour map was performed. The values obtained by this signal-matching step contained information about the presence or absence of the lesion and thus used as the decision variables. The probability distribution functions (*pdf*) of the signal-absent and signal-present decision variables were then computed. Finally, non-parametric ROC curves were derived by simple thresholding on the *pdfs* of the decision variables, and area under the ROC curves computed by the trapezoidal method.

RESULTS

Fig. 2 shows a representative case with the embedded lesion at angular projections of –22.3^o, 0^o (CC orientation), and 23.1^o. Fig. 2d shows the true positive and false positives findings of the CADe processor projected on the CC image.

Fig. 3 shows the variation in positive predictive index (PPI) with the number of projections within12 angular spans in the 3.6^o–44.8^o range. At each angular range, the PPI values first increase and then decrease with increase in the number of projections, peaking at a value that is dependent on the angular span. The maximum PPI is obtained for 10 projections spanning an angular arc of 44.8^o.

Fig. 4 shows the variation of AUC with the number of projections spanning different angular arcs. At each angular range, the AUC values increase with the increase in the number of angular projections and then appear to approach an asymptote. The number of projections at which the AUC values peak depends on the angular span. The highest AUC is obtained at an angular span of 44.8^o with 7 projections.

The trend in the variation of PPI and the AUC values delineate the role of different components of acquisition scheme in the final diagnostic performance of a multi-projection imaging system. These trends indicate that the optimum number of projections for a multi-projection imaging system may be in the 7-10 range for an angular span of 44.8^o. Most noteworthy, the observer model results (presented last year at SPIE, and reproduced here in Fig. 5) show a similar trend in performance where the maximum detectability of an embedded lesion was found to maximize with 11 angular projections for an angular span of 44.8^o.

DISCUSSION

Data acquisition parameters in multi-projection imaging modalities, such as breast tomosynthesis, are currently determined primarily by the clinical requirements such as avoiding patient motion and reducing patient discomfort.[8] However, since the acquisition scheme plays a pivotal role in the final diagnostic outcome of such a system, it is important to optimize the acquisition parameters to maximize the clinical performance of the system.

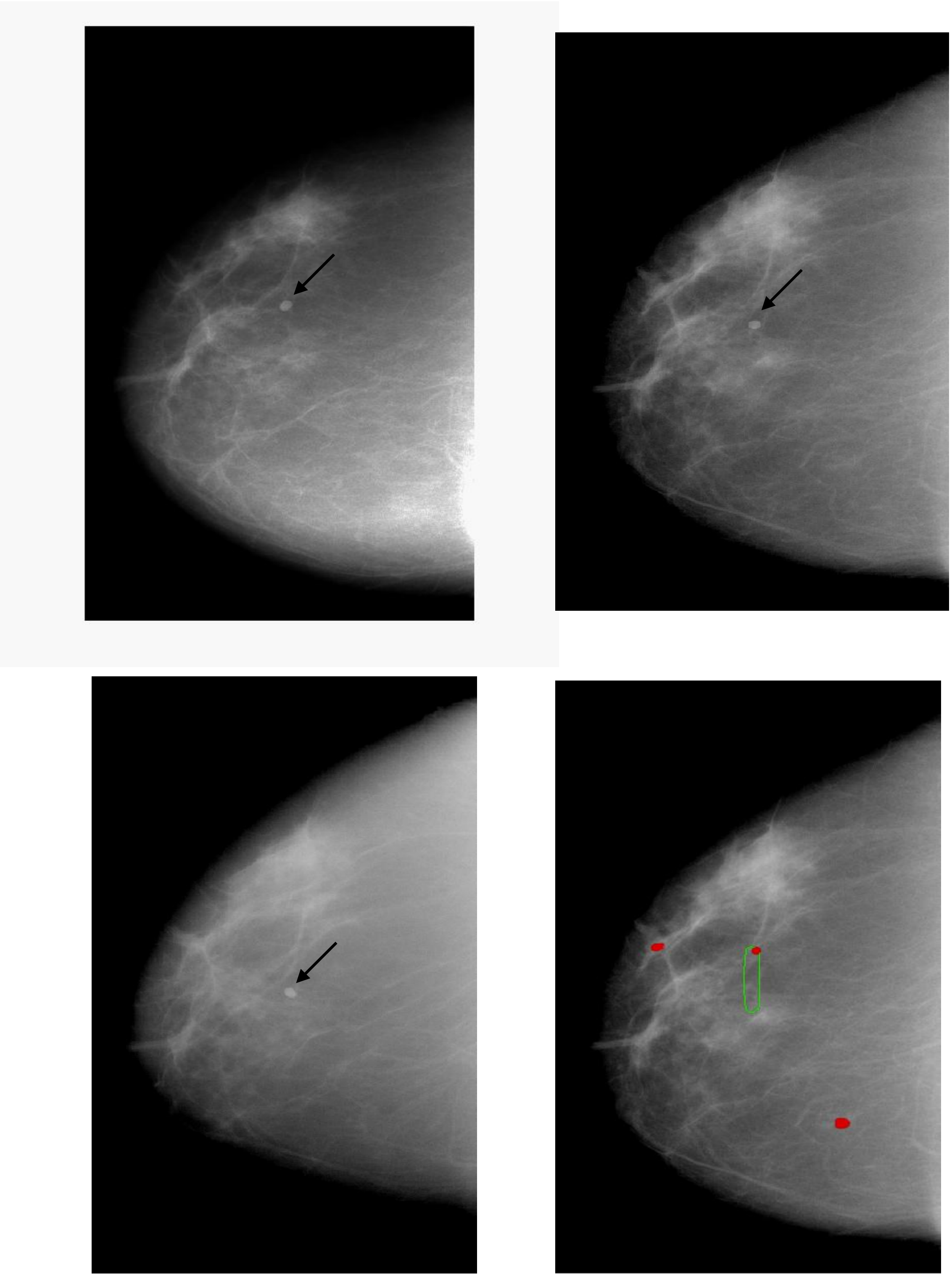
In this study, the diagnostic performance of the multi-projection imaging system was measured by two different processors that function like surrogate human observers. These were mathematical observer model-based and CADe-based processors. The two processors show how to best integrate each of the components in the acquisition scheme to maximize the performance of a multi-projection imaging system in a task that closely emulates clinical practice. While the observer model results have been reported earlier,[5] a new CADe was developed in this project to confirm the observer model results and more importantly, to study the performance of the imaging system in a setup that can potentially be implemented clincially.

The performance of the CADe processor was measured in terms of positive predictive index (PPI), which is fraction of the true positives findings to the total number of suspicious locations indicated by the CADe processor in an image set, and the area under ROC curve (AUC).

Computation of AUC was made possible because of the signal-known exactly (SKE) paradigm of our study in which a known lesion (signal) was embedded at a known location of the image. This enabled determination of the ROC curve indicative of the detectability of the embedded lesion. This technique of determining AUC to evaluate CAD does not account for the errors introduced due to search mechanism inherent to clinical diagnostic procedure, and is also not a measure of the number of false positives per image - an important consideration in benchmarking CAD performance. Nevertheless, AUC provides a robust index of metric for evaluating how effective a CAD processor is in exploiting the geometrical correlation information between multiple projections of CI for detecting a potential lesion.

At each of the 12 angular ranges considered, as the number of projections is increased, the PPI values first increase but then decrease. The increase in PPI may be attributed to the increase in true positive findings due to an increase in correlation information available from multiple projections. With further increase in projections, however, more suspicious regions come into play, thus increasing the FPs, and hence decreasing the PPI value. AUC values, on the other hand, appear to reach an asymptote beyond a certain number of projections. This is because any further increase in the number of projections offers no additional gain in the geometrical information in terms of the relative difference between the lesion and surrounding anatomical structures, thus saturating the AUC values.

Both the CADe and observer model highlighted the relationship between different components of the acquisition scheme and the diagnostic performance of a multi-projection imaging system and were successfully used to optimize the acquisition scheme to maximize the available diagnostic information in CI.



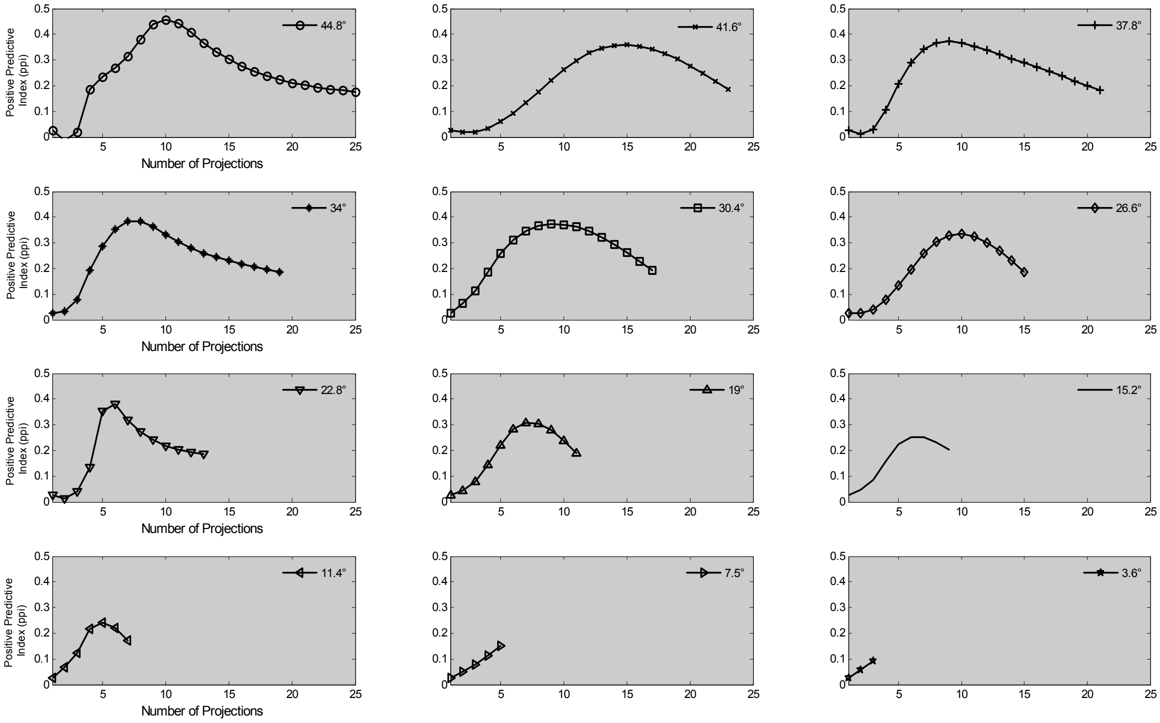


Fig. 3: Positive Predictive index [TP/(TP + FP)] as a function of the number of projections spanning different angular ranges in a multi-projection Correlation Imaging setup. TP~True Positive findings; FP~ False Positive findings per patient case.

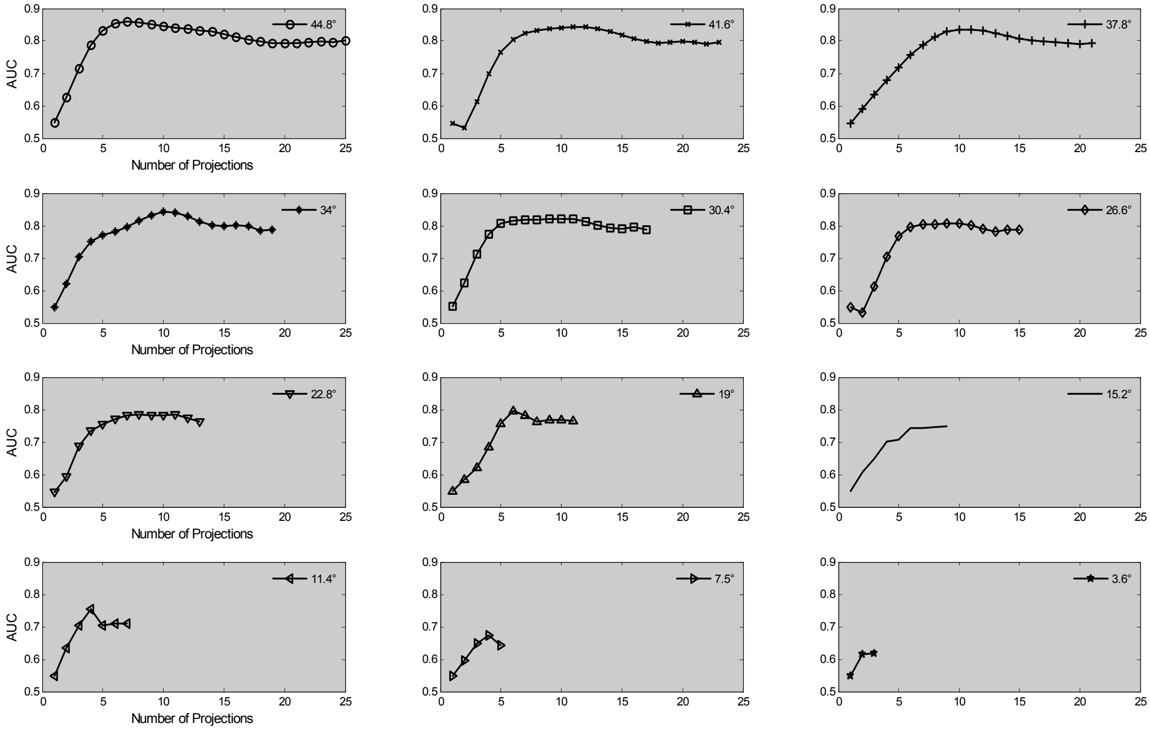


Fig. 4: Area under ROC curves as a function of the number of projections spanning different angular ranges in a multi-projection Correlation Imaging setup. AUCs indicate the detectability of a simulated mass embedded into each projection.

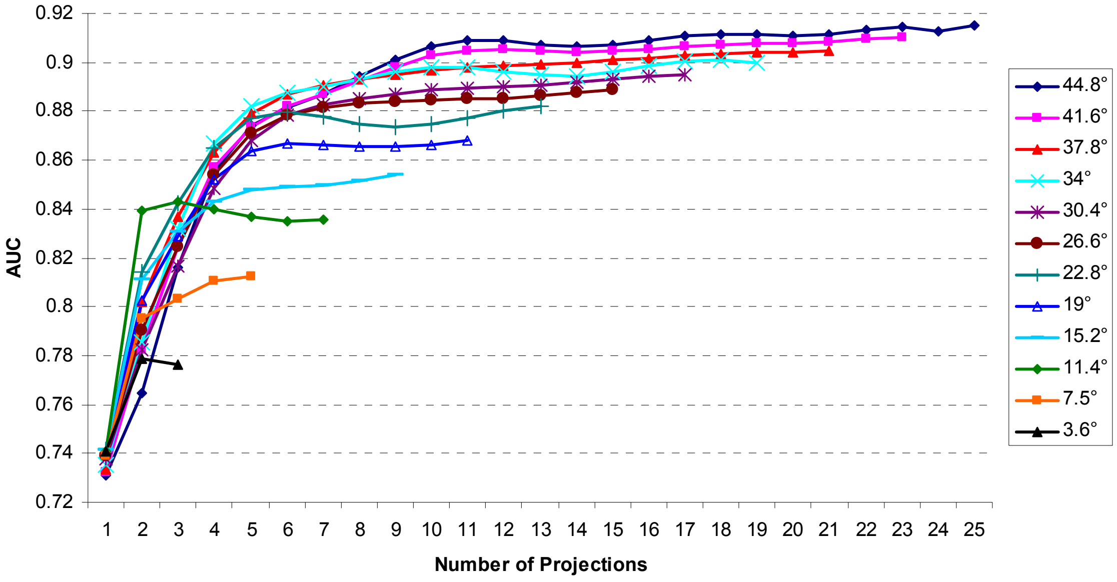


Fig. 5: Variation of AUC for different number of angular projections spanning a total angular arc in the 3.6-44.8^{o} range using a mathematical observer model.[1] These results confirm the optimization results obtained from the CADe processor.

CONCLUSIONS

A new CADe processor was developed for multi-projection Correlation Imaging (CI) that takes advantage of the geometrical correlation information to improve specificity of the CI system. The performance of the CADe system was computed at different data acquisition settings towards optimizing the geometry of image acquisition. Both the CADe and observer model results (reported earlier) show a general trend in the performance of a multi-projection imaging system as function of the different acquisition components, and confirm that the maximum performance may be obtained with 7−11 projections for an angular span of 44.8^{o}. The optimization framework presented here is generic in nature and can be used to optimize any multi-acquisition scheme, including tomosynthesis.

Acknowledgment

This work was supported in part by DOD - Pre-Doctoral Traineeship grants no. W81XWH-06-1-0449 & W81XWH-04-1-0323, and grants from Cancer Research & Prevention Foundation and the Komen foundation.

REFERENCES

[1] A. Chawla, E. Samei, R. Saunders, J. Lo and J. Baker, "A mathematical model platform for optimizing a multi-projection breast imaging system," Med. Phys. **35**, (2008).

[2] E. Samei, S. A. Stebbins, J. T. Dobbins and J. Y. Lo, "Multiprojection Correlation Imaging for Improved Detection of Pulmonary Nodules," Am. J. Roentgenol. **188**, 1239 - 1245 (2007).

[3] J. T. Dobbins and D. J. Godfrey, "Digital x-ray tomosynthesis: current state of the art and clinical potential," Phys. Med. Biol. **48**, R65-R106 (2003).

[4] A. D. A. Maidment, P. R. Bakic, M. Albert and 2003, "Effects of quantum noise and binocular summation on dose requirements in stereoradiography," Med. Phys. **30**, 3061-3071 (2003).

[5] A. Chawla, E. Samei and C. Abbey, " A mathematical model approach toward combining information from multiple image projections of the same patient," Proc. SPIE Medical Imaging **6510(1K)**, 1-11 (2007).

[6] R. S. Saunders, E. Samei, N. Majdi-Nasab and J. Y. Lo, "Initial human subject results for breast bi-plane correlation imaging technique," Proc. SPIE **6514**, 1-7 (2007).

[7] R. S. Saunders, E. Samei and C. Hoeschen, "Impact of resolution and noise characteristics of digital radiographic detectors on the detectability of lung nodules," Med. Phys. **31**, 1603-1613 (2004).

[8] T. Wu, R. H. Moore and D. B. Kopans, "Voting strategy for artifact reduction in digital breast tomosynthesis," Med. Phys. **33**, 2461 (2006).

Breast Mass Detection in Tomosynthesis Projection Images Using Information-Theoretic Similarity Measures

Swatee Singh[1,3], Georgia D. Tourassi[2,3], Joseph Y. Lo[1,2,3]

[1]Department of Biomedical Engineering
[2]Duke Medical Physics Program
[3]Duke Advanced Imaging Laboratories, Department of Radiology
Duke University Medical Center
2424 Erwin Road, Suite 302, Duke University
Durham, NC 27705
E-mail: swatee.singh@duke.edu

ABSTRACT

The purpose of this project is to study Computer Aided Detection (CADe) of breast masses for digital tomosynthesis. It is believed that tomosynthesis will show improvement over conventional mammography in detection and characterization of breast masses by removing overlapping dense fibroglandular tissue. This study used the 60 human subject cases collected as part of on-going clinical trials at Duke University. Raw projections images were used to identify suspicious regions in the algorithm's high-sensitivity, low-specificity stage using a Difference of Gaussian (DoG) filter. The filtered images were thresholded to yield initial CADe hits that were then shifted and added to yield a 3D distribution of suspicious regions. These were further summed in the depth direction to yield a flattened probability map of suspicious hits for ease of scoring. To reduce false positives, we developed an algorithm based on information theory where similarity metrics were calculated using knowledge databases consisting of tomosynthesis regions of interest (ROIs) obtained from projection images. We evaluated 5 similarity metrics to test the false positive reduction performance of our algorithm, specifically joint entropy, mutual information, Jensen difference divergence, symmetric Kullback-Liebler divergence, and conditional entropy. The best performance was achieved using the joint entropy similarity metric, resulting in ROC A_z of 0.87 ± 0.01. As a whole, the CADe system can detect breast masses in this data set with 79% sensitivity and 6.8 false positives per scan. In comparison, the original radiologists performed with only 65% sensitivity when using mammography alone, and 91% sensitivity when using tomosynthesis alone.

1. INTRODUCTION

Breast cancer is the second-most deadly type of cancer for women in the United States. The American Cancer Society estimates that 240,510 women will be diagnosed with breast cancer in 2007 alone and will kill an estimated 40,910 women. [1]. Survival rates are significantly higher when the cancer is detected at an early stage [2-4]. The 5-year survival rate (YSR) for patients with localized breast cancer is 97%. Patients with distant metastases see their 5 YSR drop to 23%. Therefore, detecting breast cancer at an early stage is critical to patient care. At present, the most common, and effective early-detection tool currently available to clinicians is screening mammography. However, mammography is well known in its inability to deal with dense overlying fibroglandular tissue. CADe was touted as a "second reader" to improve sensitivity by helping radiologists detect disease which might otherwise have been missed [4-6].

In a screening setting, radiologists typically look at 4 views per patient for mammography. However, if tomosynthesis were to replace mammography as a screening tool, then a radiologist would potentially have to look at 50 to 80 reconstructed slices per exam. This increase in the number of images will likely affect workflow dramatically. The role of CADe in such a setting becomes even more important as not just a second reader, but also to possibly identify initial suspicious breast volumes for the radiologist to focus their attention. As current

Medical Imaging 2007: Computer-Aided Diagnosis, edited by Maryellen L. Giger, Nico Karssemeijer,
Proc. of SPIE Vol. 6514, 651415, (2007) · 1605-7422/07/$18 · doi: 10.1117/12.713004

investigators in CT colonography have suggested, CADe can potentially ease radiologist workflow when working with large 3D data sets [7].

2. METHODS AND MATERIALS

2.1 Database

A prototype breast tomosynthesis system by Siemens Medical Solutions was developed. This system acquires 25 projection images over a 50-degree angular range in approximately 13 seconds. An amorphous selenium direct digital detector with a large area of 24x30 cm has been used on this system to give resultant projection images of high resolution (85 micron pixel size), which are acquired at the rate of 2 images/second. Of the over 200 human subjects that have been recruited at the Duke University Medical Center, we used 60 subjects for this study. Bilateral MLO views were acquired in screening cases, while bilateral MLO and CC views were acquired for diagnostic and biopsy cases. A single reader, Dr. Jay Baker, head of breast imaging in the department of radiology, interpreted these cases in separate and blinded readings. The gold standard was established from information available from all modalities for a patient.

Armato et al have demonstrated that the choice of a reconstruction algorithm can affect the performance of CT-CADe [8]. As such we worked with projection images rather than reconstructed slice images in order to avoid dependency on any particular reconstruction algorithm. Also, to avoid loss of inherent information present in projection images while reconstructing them, it was essential to work with projection images. Lastly, we also wished to draw upon our existing mammography-based CADe techniques, and by working with projection images - which are similar to low dose mammogram images - we were be able to achieve that objective. Other groups have also implemented reconstruction independent CAD [9-10].

2.2 Experimental Design

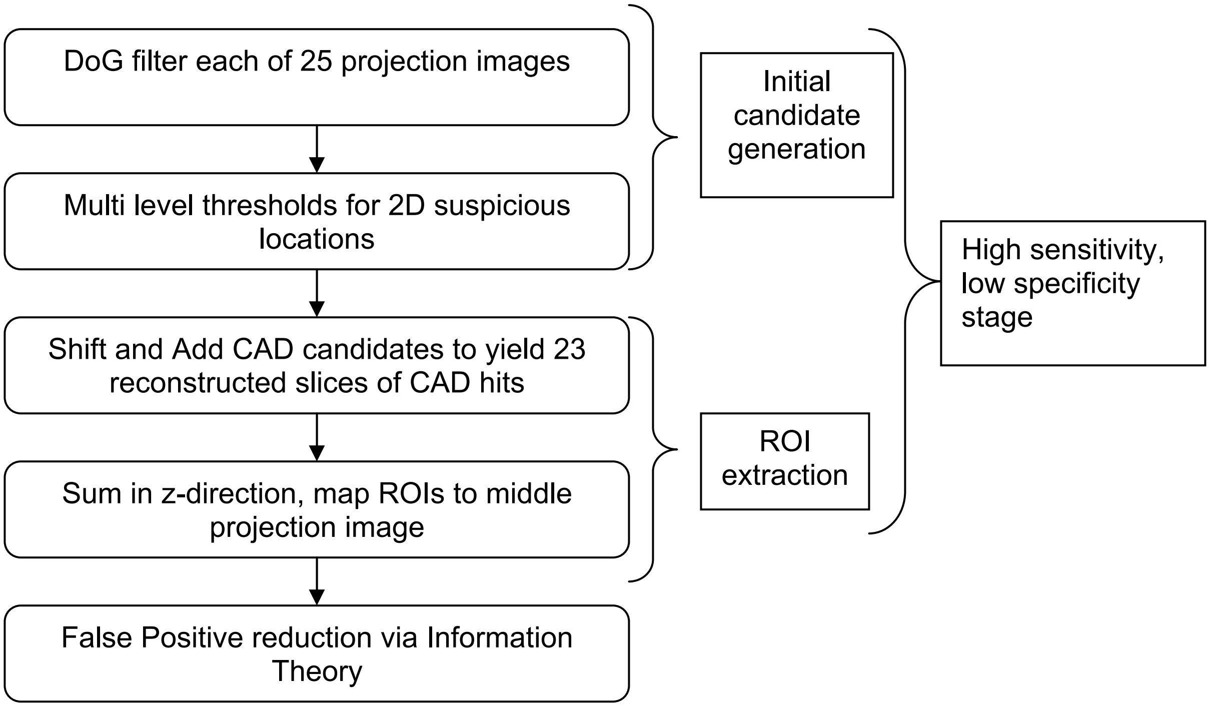
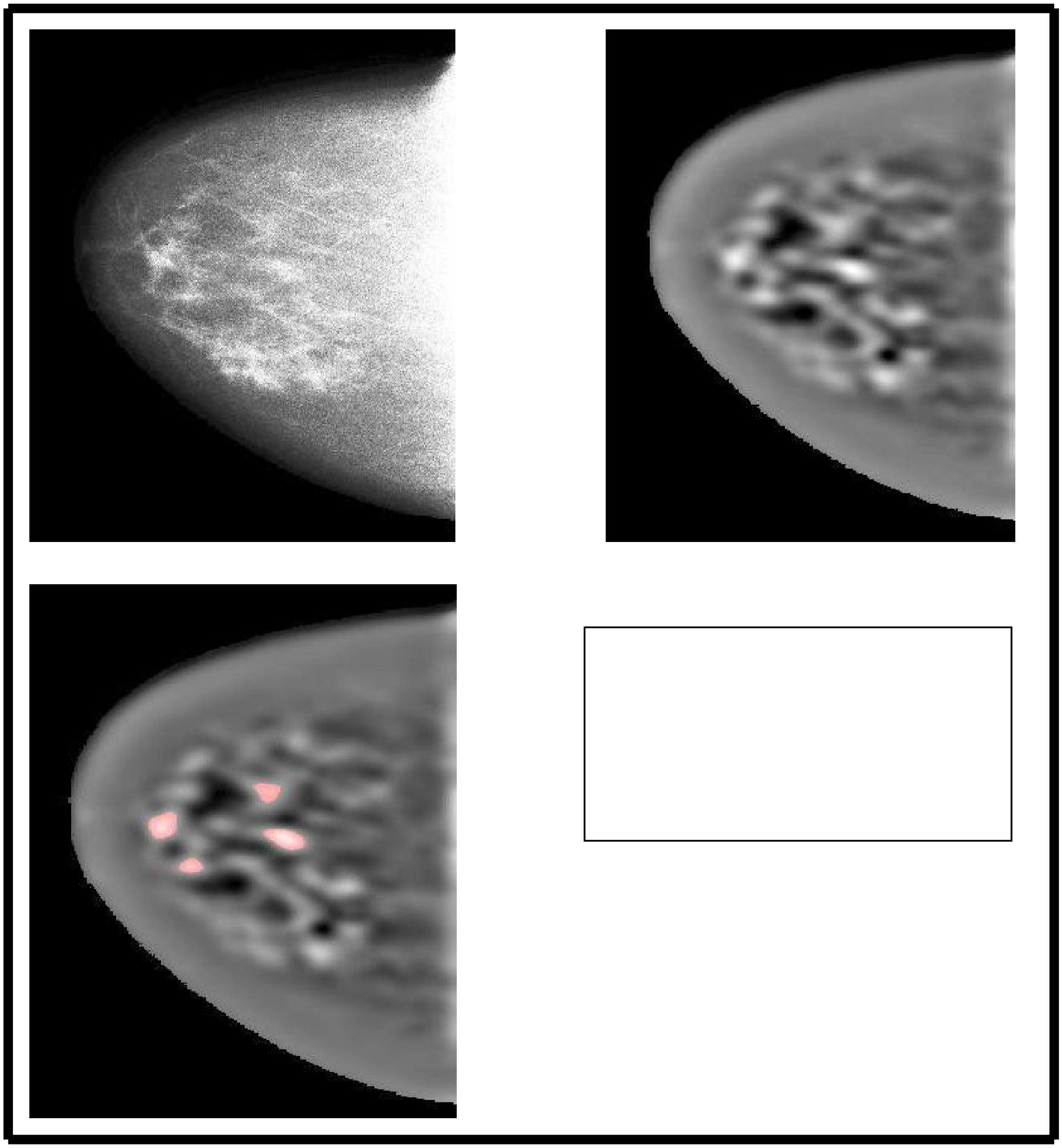

DoG filter each of 25 projection images
Multi level thresholds for 2D suspicious locations
Shift and Add CAD candidates to yield 23 reconstructed slices of CAD hits
Sum in z-direction, map ROIs to middle projection image
False Positive reduction via Information Theory
Initial candidate generation
ROI extraction
High sensitivity, low specificity stage

Figure 1: The algorithm adopted in this study for our CADe system



2.3 High-sensitivity, low-specificity stage

We have included in this paper, as an example subject 33's LCC scan's middle projection to visually demonstrate the high-sensitivity, low-specificity stage of the algorithm. Figure 2(a) shows the middle projection of the scan for subject 33. We DoG filter this projection image to yield an image shown in figure 2(b). This image is then subjected to multi-level thresholding to yield 2-D CADe suspicious locations. Figure 2(c) shows the shift-and-added 2-D CAD suspicious locations in red overlaid on the filtered image.

2.4 False Positive Reduction

To reduce the number of false positives, we used information theoretic principles to assess image similarity.

Given a query tomosynthesis ROI *Qi* , a decision index *D*(*Qi*) was calculated as the difference of two terms. Assuming that the knowledge database contains *k* mass cases and *l* normal cases. The first term of the decision index *D*(*Qi*) measures the average MI between the query ROI and its *k* best mass matches M_j . Similarly, the second term measures the average MI between the query ROI and its *l* best normal N_j matches,

$$D(Q_i) = \frac{1}{k}\sum_{j=1}^{k} MI(Q_i, M_j) - \frac{1}{l}\sum_{j=1}^{l} MI(Q_i, N_j)$$

Theoretically, a query ROI depicting a mass should have a higher *D*(*Qi*). For this study 60 cases were used. This algorithm’s initial high sensitivity, low specificity stage yielded ROIs that were extracted from the middle projection. Hence, false positive reduction was done using only ROIs obtained from the middle projection image of each scan. Results were reported as Receiver Operating Characteristic (ROC) Area Under Curve (AUC) by applying a leave-one-out cross validation scheme on all available ROIs.

2.4 Similarity Metrics

For this experiment, we measured five similarity metrics: (1) joint entropy, (2) average conditional entropy, (3) mutual information, (4) maximum Kullback-Leibler divergence and, (5) Jensen divergence. These metrics were measured as follows:

1. Joint entropy: $Jo\text{int}\, H = -\sum_x \sum_y p_{XY}(x,y)\log\left(p_{XY}(x,y)\right)$

2. Average conditional entropy: $\overline{conditional\_H} = \frac{H(x\,|\,y) + H(y\,|\,x)}{2}$

3. Mutual Information: $MI(X,Y) = \sum_x \sum_y P_{XY}(X,Y)\log_2\left(\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right)$

4. Maximum Kullback-Leibler divergence: $\max\_divergence = \max\left(D(q\,||\,p), D(p\,||\,q)\right)$

where, Kullback-Leibler divergence: $avg\_divergence = \frac{D(q\,||\,p) + D(p\,||\,q)}{2}$

and $D(q\,||\,p) = \sum_x q(x)\log\left(\frac{q(x)}{p(x)}\right)$

5. Jensen divergence: $JD(p,q)=\sum_{x}\left(q(x)\log\frac{2q(x)}{p(x)+q(x)}+p(x)\log\frac{2p(x)}{p(x)+q(x)}\right)$

For further details about these measures and their specific application in our system, please refer to the Medical Physics paper by co-author Georgia Tourassi [11].

3. RESULTS

Results from the first stage of the algorithm, the high sensitivity low specificity stage, are reported in the form of a Free-response Receiver Operating Characteristic (FROC) curve in figure 4 below.

3.1 Performance of Measures Similarity Metrics

Table 1 below lists the individual performance of the 5 different similarity metrics used in this study.

Similarity Metric	**ROC A_z**
joint entropy	0.87 ± 0.01
mutual information	0.83 ± 0.01
conditional entropy	0.83 ± 0.01
maximum symmetric Kullback-Leibler divergence	0.77 ± 0.02
Jensen difference divergence	0.72 ± 0.04

Table 1: Performance of various similarity metrics for false positive reduction

Our best performing metric was joint entropy with an A_z of 0.87 and has been individually plotted in figure 3.

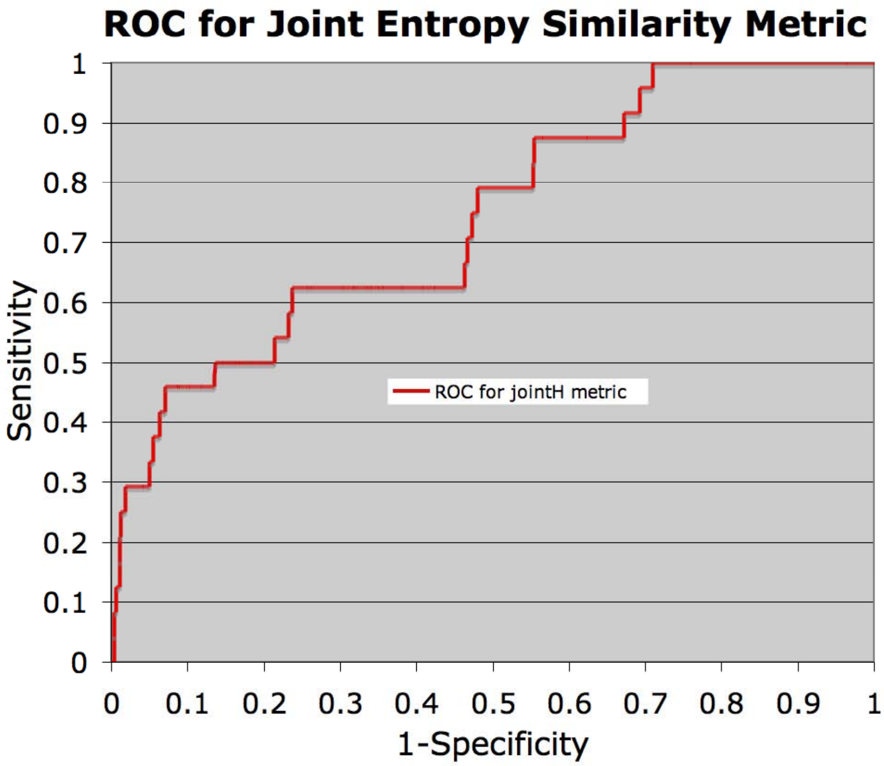


Figure 3: ROC of the best performing similarity metric

3.2 Overall System Performance

In the area where we care about the most - where sensitivity is around 90% - all 5 metrics perform about the same. As such, we picked our best performing metric, the joint entropy and chose to work at an operating point where the maximum sensitivity was 87.5% with a 45% reduction in false positives from the initial stage of the algorithm. This gives us an overall system performance. Figure 4 shows two Free-response Receiver Operating Characteristic (FROC) curves. The line in blue is the FROC curve for the initial, high-sensitivity, low-specificity stage of the algorithm. After we apply the operating point picked for the joint entropy metric to this FROC, we get an overall system FROC curve plotted in red in figure 4. The maximum sensitivity of the overall system is now 79% with 6.8 false positives / scan.

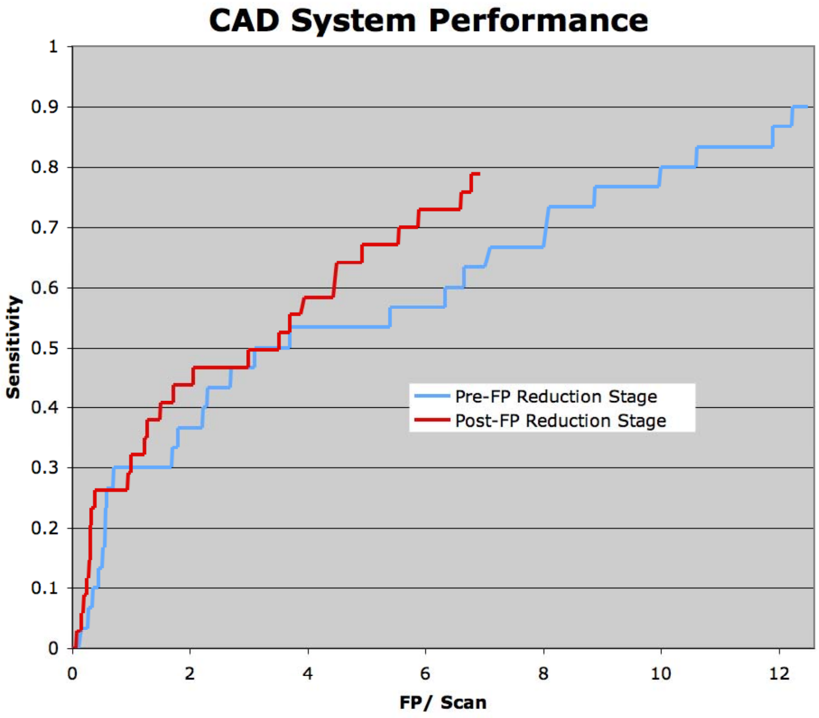


Figure 4: Overall system performance. The blue line depicts the FROC for the pre-false positive reduction stage, while the red line shows the overall performance of the CADe system after the false positive reduction stage

3.3 Human Subject Example

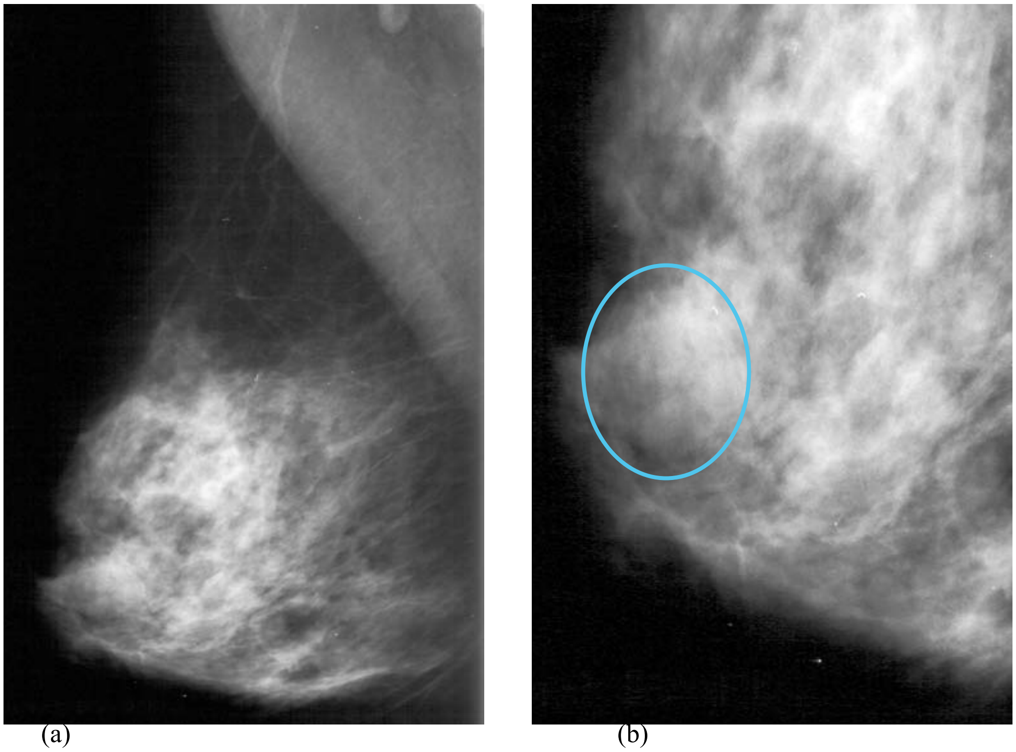
As an example, we picked subject 60 that has a lesion that was not detected with mammography. Figure 5(a) below shows a scanned film of the RMLO view. Also in figure 5(b) we have shown a zoomed out version of that breast where the lesion is actually present.

In figure 6(a) we show the tomosynthesis reconstructed slice number 15 of this breast view with the cancer clearly visible and encircled in red for ease. Figure 6(b) is the same reconstructed slice with the CADe hits overlaid on it as a cross. As we can see, the lesion was correctly picked out by the CADe algorithm along with one false positive.

4. CONCLUSIONS

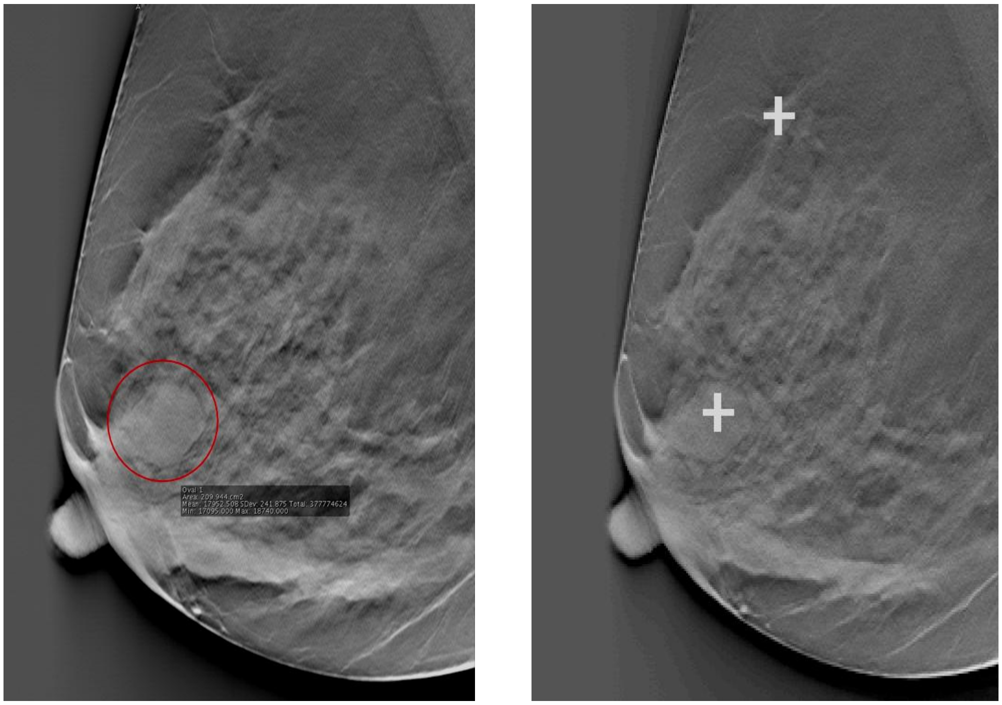
The role of CADe is especially important in breast tomosynthesis due to the large volume of data. If this modality is ever intended as a screening tool, then a CADe algorithm that presents the radiologist with initial cues could potentially become indispensable for reading large patient data in a reasonable amount of time.

Despite a difficult dataset where the radiologist performance was only 65% using mammograms alone, our CADe algorithm achieved a maximum sensitivity of 79% on a per scan basis with a maximum false positive per scan rate of 6.8. Also, we saw encouraging performance of the information theory false positive reduction stage despite a



(a) (b)

Figure 5: (a) mammographic film of subject 60 (b) magnified region of interest of subject 60 with the obscured lesion encircled in blue


Oval 1
Area: 209.944 cm2
Mean: 17952.508 SDev: 241.875 Total: 377774624
Min: 17095.000 Max: 18740.000

(a) (b)

Figure 6: (a) Reconstructed slice number 15 of subject 60 with the lesion encircled in red (b) Reconstructed slice number 15 of subject 60 with the CADe hits for that slice overlaid on it as white crosses.

small mass database to serve as the knowledge database. Thus, we have demonstrated feasibility of developing a CADe algorithm using models with the extremely low-dose tomosynthesis projection slices. Future work will expand the data set size, explore direct optimization of the CADe techniques for tomosynthesis projection images, and increase the size of the knowledge database to improve performance of the false positive reduction stage of the algorithm.

ACKNOWLEDGMENTS

This work was supported in part by US Army Breast Cancer Research Program W81XWH-05-1-0293, NIH/NCI R01 CA101911 and R01 CA112437, and Siemens Medical Solutions.

REFERENCES

[1] ACS, "American Cancer Society: Cancer Facts and Figures 2004. Atlanta, Ga: American Cancer Society 2004.," 2004.

[2] I. Anttinen, M. Pamilo, M. Soiva, and M. Roiha, "Double reading of mammography screening films: one radiologist or two?" *Clinical Radiology*, vol. 48, pp. 414-421, 1993.

[3] W. R. Hendee, C. Beam, and E. Hendrick, "Proposition: all mammograms should be double-read," *Medical Physics*, vol. 26, pp. 115-118, 1999.

[4] E. L. Thurfjell, K. A. Lernevall, and A. A. S. Taube, "Benefit of independent double reading in a population-based mammography screening program," *Radiology*, vol. 191, pp. 241-244, 1994.

[5] T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center," *Radiology*, vol. 220, pp. 781-786, 2001.

[6] C. E. Metz and J. H. Shen, "Gains in accuracy from replicated readings of diagnostic images," *Medical Decision Making*, vol. 12, pp. 60-75, 1992.

[7] A.H. Dachman, and H. Yoshida, "Virtual Colonoscopy: past, present, and future," *Radiol Clin North Am*, 41(2), pp377-93, 2003.

[8] S. G. Armato III, M. B. Altman, and P. J. La Riviere, "Automated detection of lung nodules in CT scans: effect of image reconstruction algorithm," *Medical Physics*, vol. 30, pp. 461- 472, 2003.

[9] G. Peters, S. Muller, S. Bernard, R. Iordache, and I. Bloch, "Reconstruction-independent 3D CAD for mass detection in digital breast tomosynthesis using fuzzy particles," *Proc. SPIE*, 2006, vol. 6144, pp. 661-670, 2006.

[10] I. Reiser, R. M. Nishikawa, M. L. Giger, T. Wu, E. A. Rafferty, R. Moore, and D. B. Kopans, "Computerized mass detection for digitial breast tomosynthesis directly from the projection images," *Medical Physics*, vol. 33, pp. 482-491, 2006.

[11] G. D. Tourassi, B. Harrawood, S. Singh, J. Y. Lo, and C. E. Floyd, “Evaluation of information theoretic similarity measures for content-based retrieval and detection of masses in mammograms,” *Medical Physics*, vol. 34, pp.140-150, 2007.

Feasibility Study of Breast Tomosynthesis CAD System

Anna Jerebko[1], Yuan Quan[1], Nicolas Merlet[2], Eli Ratner[2],
Swatee Singh[3], Joseph Y. Lo[3], Arun Krishnan[1]
[1]Siemens Med., Malvern PA, USA,
[2]Siemens Med., Jerusalem, Israel,
[3]Duke University, NC, USA

ABSTRACT

The purpose of this study was to investigate feasibility of computer-aided detection of masses and calcification clusters in breast tomosynthesis images and obtain reliable estimates of sensitivity and false positive rate on an independent test set. Automatic mass and calcification detection algorithms developed for film and digital mammography images were applied without any adaptation or retraining to tomosynthesis projection images. Test set contained 36 patients including 16 patients with 20 known malignant lesions, 4 of which were missed by the radiologists in conventional mammography images and found only in retrospect in tomosynthesis. Median filter was applied to tomosynthesis projection images. Detection algorithm yielded 80% sensitivity and 5.3 false positives per breast for calcification and mass detection algorithms combined. Out of 4 masses missed by radiologists in conventional mammography images, 2 were found by the mass detection algorithm in tomosynthesis images.

Keywords: computer-aided diagnosis, mammography, tomosynthesis.

1. INTRODUCTION

Breast tomosynthesis is an investigational 3D imaging technique with the potential to improve sensitivity and specificity of breast cancer diagnosis. The technique can produce high resolution reconstructed slice images with time and dose comparable to conventional mammography. However, the large number of images produced can dramatically impact radiologist workflow and lead to missed lesions due to fatigue. Therefore, computer-aided detection (CAD) algorithms may play an integral role in the use of breast tomosynthesis, even more so than they already do in conventional mammography [1]-[7].

1.1. Literature review

In the past 3 years several research groups published very promising results on breast tomosynthesis CAD, including detection of masses and calcification clusters in reconstructed 3D images as well as in original projections. Chan et al. in [8] proposed a gradient field based algorithm for breast mass detection in reconstructed 3D tomosynthesis image. The algorithm's performance was evaluated on 26 patients (23 masses) by means of leave-one-out (LOO) validation method as 85% sensitivity with 2.2 false-positives per case. The same authors recently published the results [9] with sensitivity of 90% and false positives rate reduced to 1.2 per case evaluated on the same data set.

Several papers suggest using original projection images for calcification and mass detection and using the 3D reconstructed image only in the final stage. Wheeler et al. [10] presented an algorithm for calcification detection in the projection images. Calcification residual images are then reconstructed into a 3D image and final assessment is made in 3D space. Sensitivity and false positive rate are not specified in this paper.

Peters et al. [11] suggested a mass detection algorithm where 3D reconstruction is also used in the final stage only. The sensitivity of this algorithm was 86%, but only 7 masses were available in the test set. The false positive rate was 3.5 per case, evaluated on 4 normal cases.

Medical Imaging 2007: Computer-Aided Diagnosis, edited by Maryellen L. Giger, Nico Karssemeijer,
Proc. of SPIE Vol. 6514, 651414, (2007) · 1605-7422/07/$18 · doi: 10.1117/12.712729

Similar workflow is used by Reiser et al. [12]. The original projections are used for primary lesion (mass) detection and then "the locations of a lesion candidate are backprojected" into a 3D reconstruction image. The algorithm's performance (sensitivity of 90% at 1.5 false positives per breast) was evaluated on the same data set which was used for development: 21 cases with masses (13 malignant) and 15 normal cases. The number of false positives was significantly reduced compared with the previous paper by the same authors [13] where it was 13 FPs per case with the same sensitivity. As the authors state in their latest paper [12], the sensitivity and false positive estimates could be positively biased due to the lack of data and absence of an independent test set.

This statement will likely hold true for all of the algorithm descriptions and evaluations mentioned above [8]-[13]: it is always hard to predict generalization properties of a CAD algorithm which was developed and tested using the same extremely limited data set. It is also true that in medical imaging in general it is hard to collect an amount data with consistent acquisition protocol that is enough to obtain statistically significant estimates of sensitivity and specificity. This becomes apparent especially in the new emerging modalities like tomosynthesis, where acquisition protocols, number of projections, and reconstruction algorithms vary significantly for different research centers. Clearly, while there is an interest among clinicians and computer scientists in developing CAD algorithms for breast tomosynthesis, much more extensive evaluation on independent data sets is needed to establish CAD feasibility for this new quickly developing modality.

The purpose of this study was to establish a reliable base-line mass and calcification detection algorithm performance assessment on an independent test set by investigating whether an existing CAD algorithm designed for mammography [1]-[7] can be applied without any re-training to breast tomosynthesis images.

2. METHODS

A prototype breast tomosynthesis system by Siemens Medical Solutions was developed to acquire 25 projection images over a 50 degree angular range in approximately 13 seconds. The system uses an amorphous selenium direct digital detector with a large area (23x30 cm), high resolution (85 micron pixel size), and 2 images/second frame rate. The system has been undergoing evaluation and clinical trials at Duke University Medical Center. One hundred human subjects were recruited at that site, consisting of 65 routine screening, 25 diagnostic mammography, and 10 cases undergoing biopsy. This study used the first 100 human subject cases collected as part of on-going clinical trials.

2.1. Mammo CAD algorithm

Our computer aided detection algorithm included three major stages: candidate generation, feature extraction and classification. Two separate algorithms were designed for detection of calcification clusters and masses. Both algorithms employed multiple features characterizing various aspects of density, texture, shape and size of potential (candidate) findings and healthy tissue around them. In addition to a cascade of filters aimed at reducing the number of candidates in the process of candidate generation and feature extraction, the final classification scheme was applied to classify the lesions according to their likelihood of malignancy.

The features and classification scheme were originally developed for regular mammograms. The classification scheme was constructed using a training procedure on a separate dataset of 553 regular screen-film mammogram cases from various BI-RADS categories, with proven pathology (281 malignant), including 200 mass lesions and 353 clusters. Adaptations were made in the image preprocessing step for FFDM (digital mammography) images.

2.2. Mass and calcification detection in tomosynthesis images

The only adaptation of Mammo CAD algorithm for low dose tomosynthesis images involved median filtering to reduce noise. Then the mass and calcification detection algorithms were applied without any changes. The outline of the algorithm and the dataflow are schematically shown in figure 1 below.

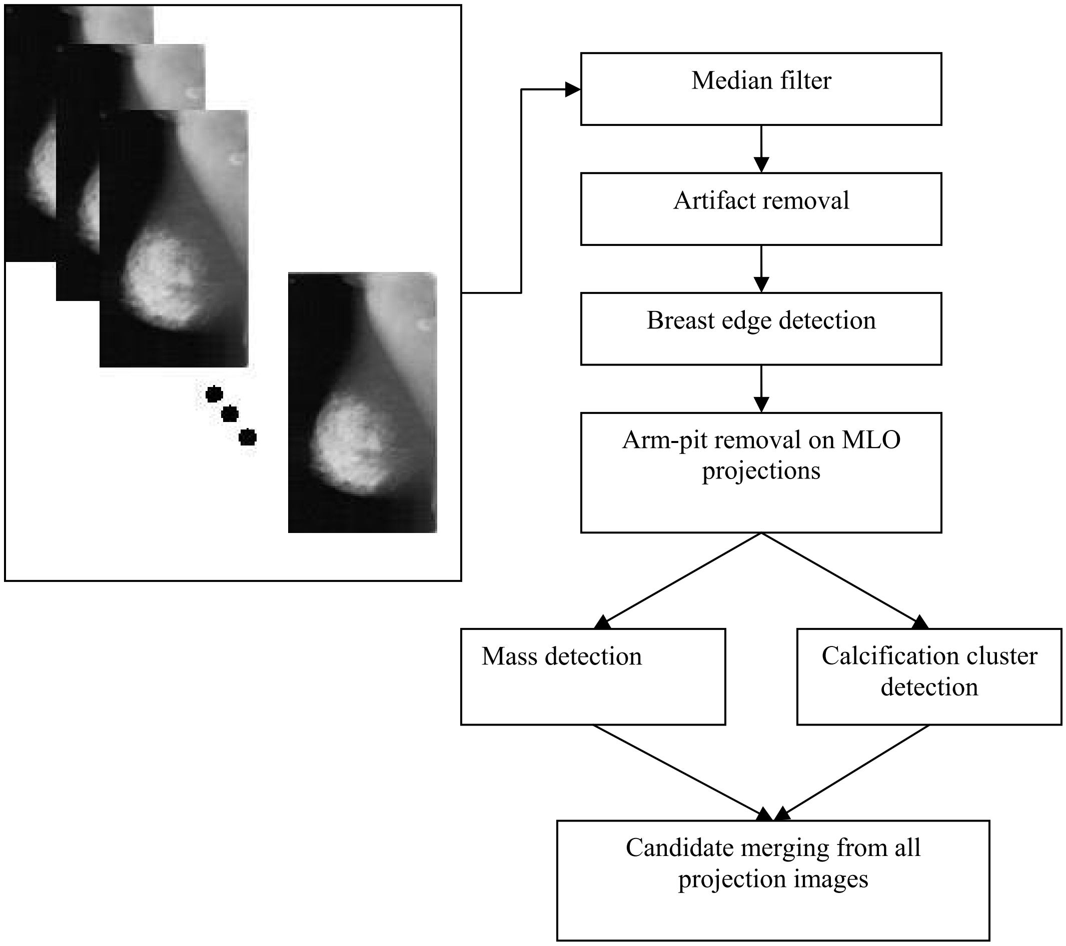

Median filter
Artifact removal
Breast edge detection
Arm-pit removal on MLO projections
Mass detection
Calcification cluster detection
Candidate merging from all projection images

Fig. 1. Outline of mass and calcification cluster detection algorithm for tomosynthesis images.

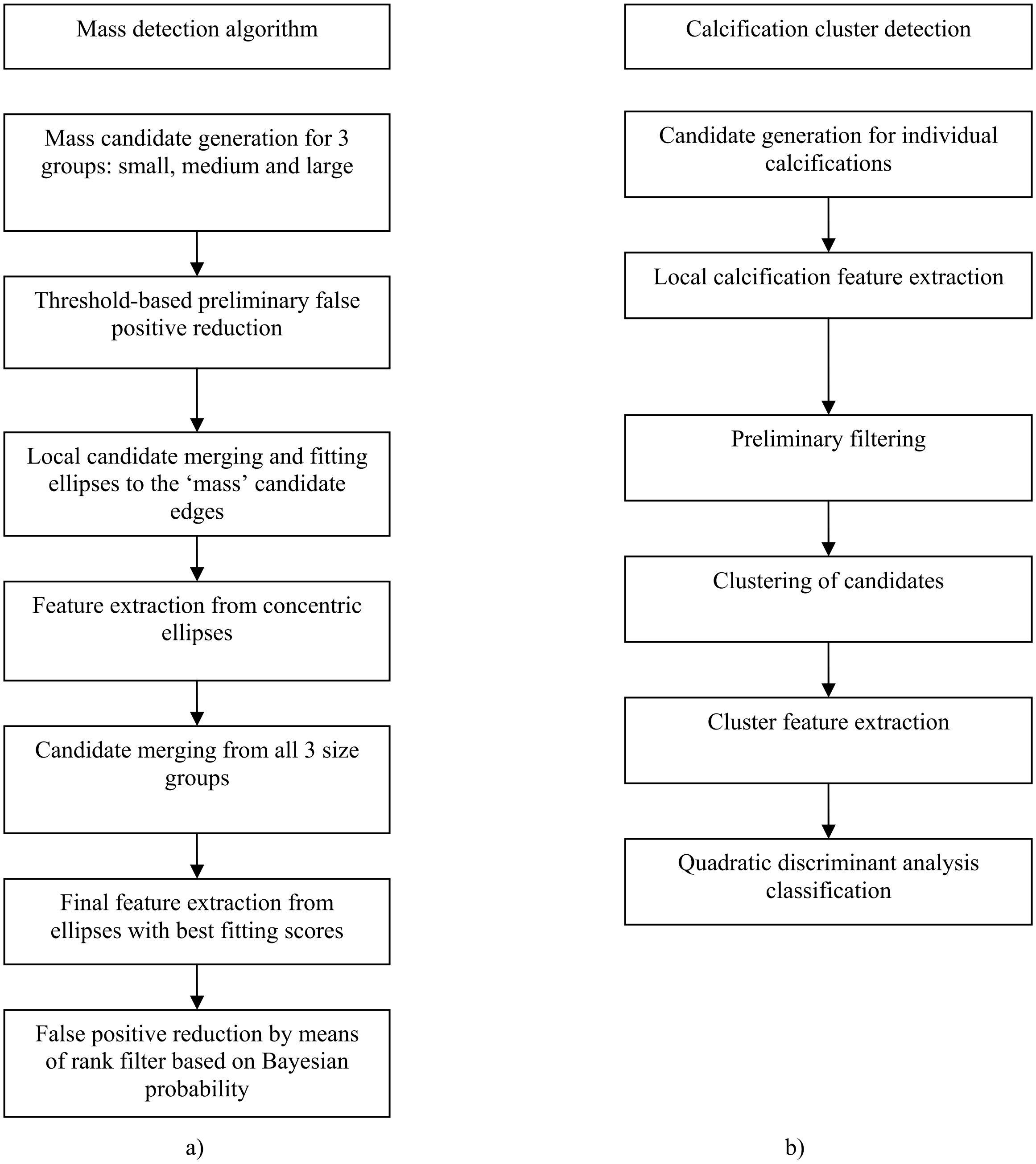
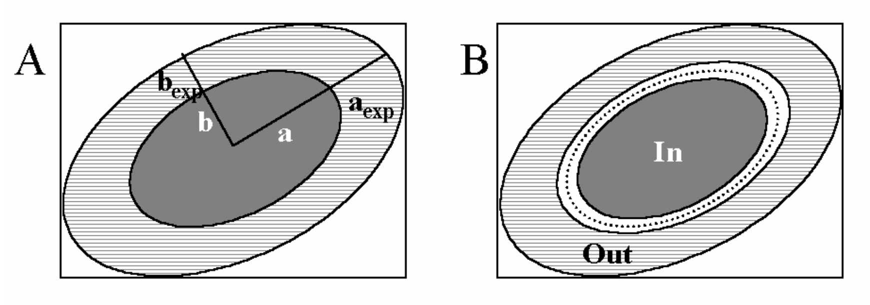

Mass detection algorithm
Mass candidate generation for 3 groups: small, medium and large
Threshold-based preliminary false positive reduction
Local candidate merging and fitting ellipses to the 'mass' candidate edges
Feature extraction from concentric ellipses
Candidate merging from all 3 size groups
Final feature extraction from ellipses with best fitting scores
False positive reduction by means of rank filter based on Bayesian probability
a)
Calcification cluster detection
Candidate generation for individual calcifications
Local calcification feature extraction
Preliminary filtering
Clustering of candidates
Cluster feature extraction
Quadratic discriminant analysis classification
b)

Fig. 2. Outline of CAD algorithm developed based on mammography images.
a) Mass detection algorithm. b) Calcification detection algorithm.


A
b_{exp}
b
a
a_{exp}
B
In
Out

$$Str = \frac{\sum \left(g_i \cdot krn_i\right)}{\sqrt{\sum g_{i}^{2}} \cdot \sqrt{\sum krn_i^2}} \tag{4}$$

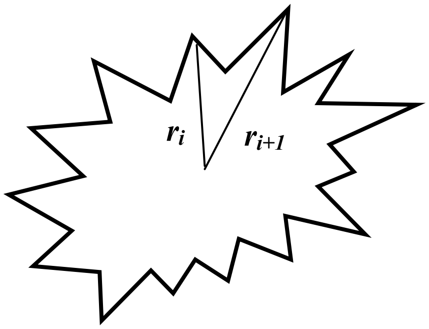


Fig. 4. Mass border irregularity evaluation.

Mass spiculation and border irregularity Jbw was evaluated through relationship between the average of absolute values of differences between two neighbor radii of polygon constructed from mass border points and the average radius of the polygon as follows:

$$\begin{gathered} \Delta r_i = \left| r_{i+1} - r_i \right| \\ Jbw = \frac{\sum \Delta r_i}{\sum r_i} \end{gathered} \tag{5}$$

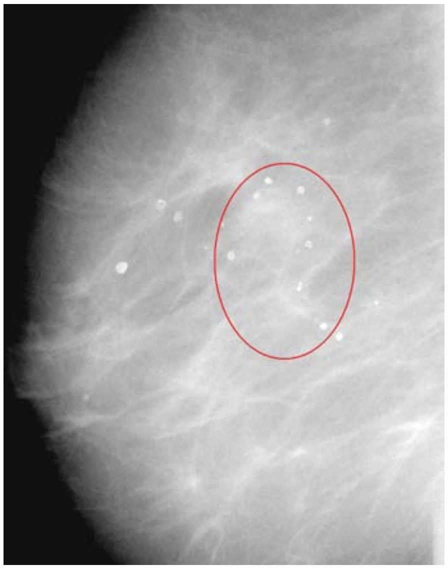
After threshold-based preliminary false positive reduction, candidates from all three size groups were merged. Final ellipse fitting and additional feature extraction are followed by false positive reduction by means of rank filter based on Bayesian probability.

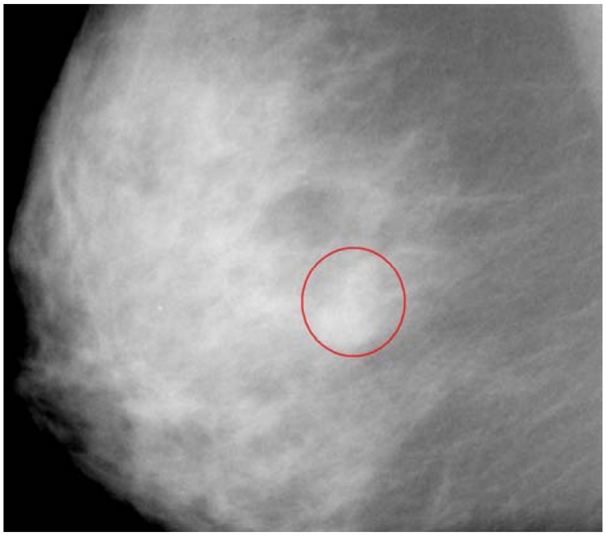
2.4. Calcification cluster detection.

Candidate generation aimed at finding local intensity peaks was used as a first step to detect microcalcifications. After the candidate merging and clustering procedure, features were computed to characterize density, shape and morphology of clusters as well as individual calcifications. In addition to texture and intensity features, features capturing cluster morphology and shape were designed [1],[3],[5],[7]. Morphology was analyzed via average number of neighbors for each calcification in Delaunay triangulation representation of the cluster. Cluster shape (round or elongated) was reflected in the eccentricity feature. To characterize the shape of individual calcifications within the clusters, the ratio of the area of the microcalcification to its radius was computed and the average value assigned to the whole cluster. Finally stepwise feature selection and quadratic discriminant analysis classification were applied for false positive reduction.

2.5. Testing on tomosynthesis images

No changes were made to any of the stages of the algorithm, including the classifier. The test set consisted of tomosynthesis images of 36 patients including 16 patients with 20 known malignant lesions 4 of which were missed by the radiologists in conventional mammography images and found only in retrospect in tomosynthesis. Two views (CC and MLO) per breast, each containing 25 projection images were considered. The original low-dose tomosynthesis projection images were preprocessed using only a median filter to reduce noise.





provided that these rates are obtained from an independent test set. Future work will expand the data set size, explore direct optimization of the CAD techniques for tomosynthesis projection as well as 3D reconstructed images, and investigate other reconstruction algorithms.

REFERENCES

1. I. Leichter, R. Lederman, S.S. Buchbinder, P. Bamberger, B. Novak, S. Fields. Computerized evaluation of mammographic lesions: what diagnostic role does the shape of the individual microcalcifications play compared with the geometry of the cluster? AJR Am J Roentgenol. 2004 Mar;182(3):705-12.
2. S.S. Buchbinder, I. Leichter, R. Lederman, B. Novak, P. Bamberger, M. Sklair-Levy, G. Yarmish, S. Fields. Computer-aided classification of BI-RADS category 3 breast lesions. Radiology. 2004 Mar;230(3):820-3.
3. S.S. Buchbinder, I. Leichter, R. Lederman, B. Novak, P. Bamberger, H. Coopersmith, S. Fields. Can the size of microcalcifications predict malignancy of clusters at mammography? Acad Radiol. 2002 Jan;9(1):18-25.
4. I. Leichter, S. Buchbinder, P. Bamberger, B. Novak, S. Fields, R. Lederman. Quantitative characterization of mass lesions on digitized mammograms for computer-assisted diagnosis. Invest Radiol. 2000 Jun;35(6):366-72.
5. I. Leichter, R. Lederman, S. Buchbinder, P. Bamberger, B. Novak, S. Fields. Optimizing parameters for computer-aided diagnosis of microcalcifications at mammography. Acad Radiol. 2000 Jun;7(6):406-12.
6. I. Leichter, S. Fields, R. Nirel, P. Bamberger, B. Novak, R. Lederman, S. Buchbinder. Improved mammographic interpretation of masses using computer-aided diagnosis. Eur Radiol. 2000;10(2):377-83.
7. I. Leichter, R. Lederman, P. Bamberger, B. Novak, S. Fields, S.S. Buchbinder. The use of an interactive software program for quantitative characterization of microcalcifications on digitized film-screen mammograms. Invest Radiol. 1999 Jun; 34(6):394-400.
8. H.-P.Chan, J. Wei, B. Sahiner, E. A. Rafferty, T. Wu, M. A. Roubidoux, R. H. Moore, D. B. Kopans, L. M. Hadjiiski, and M. A. Helvie. Computer-aided Detection System for Breast Masses on Digital Tomosynthesis Mammograms: Preliminary Experience. *Radiology* 2005;237:1075-1080.
9. H.-P.Chan, J. Wei, M.A. Helvie. R.H. Moore, D.B. Kopans, Digital Breast Tomosynthesis (DBT) Mammography: Computer-aided Mass Detection by Fusion of Tomosynthesis and 3D Mass Likelihood Information. RSNA 2006
10. F. W. Wheeler, A. G. Amitha Perera, B. E. Claus, S. L. Muller, G. Peters, J. P. Kaufhold. Micro-calcification detection in digital tomosynthesis mammography Proc. SPIE Vol. 6144, 614420, Medical Imaging 2006: Image Processing; Joseph M. Reinhardt, Josien P. Pluim; Eds. Mar 2006
11. G. Peters, S. Muller, S. Bernard, R. Iordache, I. Bloch. Reconstruction-independent 3D CAD for mass detection in digital breast tomosynthesis using fuzzy particles Proc. SPIE Vol. 6144, 61441Z, Medical Imaging 2006.
12. I. Reiser, R. M. Nishikawa, and M. L. Giger T. Wu, E. A. Rafferty, R. Moore, and D. B. Kopans. Computerized mass detection for digital breast tomosynthesis directly from the projection images. Medical Physics. 2006 Feb;33(2):482-91.
13. I.Reiser, R.M. Nishikawa, M.L. Giger, D.B. Kopans, E.A. Rafferty, T. Wu and R. Moore. A multi-scale 3D radial gradient filter for computerized mass detection in digital tomosynthesis breast images. CARS 2005. 1058-1063